Cheminformatics Approaches to Structure Based Virtual Screening:

Methodology Development and Applications

Jui-Hua Hsieh

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Pharmacy (Division of Medicinal Chemistry and Natural Products)

Chapel Hill
2011

Approved by

Dr. Alexander Tropsha

Dr. Michael Jarstfer

Dr. Stephen Frye

Dr. Nikolay Dokholyan

Dr. Scott Singleton

Abstract

Jui-Hua Hsieh: Cheminformatics Approaches to Structure Based Virtual Screening: Methodology Development and Applications
(Under the direction of **Dr. Alexander Tropsha)**

Structure-based virtual screening (VS) using 3D structures of protein targets has become a popular *in silico* drug discovery approach. The success of VS relies on the quality of underlying scoring functions. Despite of the success of structure-based VS in several reported cases, target-dependent VS performance and poor binding affinity predictions are well-known drawbacks in structure-based scoring functions. The goal of my dissertation is to use cheminformatics approaches to address above problems of the existing structure-based scoring methods.

In Aim 1, cheminformatics practices are applied to those problems which conventional structure-based scoring functions find difficult (anti-bacterial leads efflux study) or fail to address (AmpC β-lactamase study). Predictive binary classification QSAR models can be constructed to classify complex efflux properties (low *vs.* high) and to differentiate AmpC β-lactamase binders from binding decoys (i.e., the false positives generated by scoring functions). The above models are applied to virtual screening and many computational hits are experimentally confirmed.

In Aim 2, novel statistical *binding* and *pose* scoring functions (or pose filter in Aim 3) are developed, to accurately predict protein-ligand binding affinity and to discriminate native-like poses of ligands from pose decoys respectively. In my approach, the protein-ligand interface is represented at the atomic level resolution and transformed via a special

computational geometry approach called Delaunay tessellation to a collection of atom quadruplet motifs. And individual atom members of the motifs are characterized by conceptual Density Functional Theory (DFT)-based atomic properties. The *binding* scoring function shows acceptable prediction accuracy towards Community Structure-Activity Resources (CSAR) data sets with diverse protein families.

In Aim 3, a two-step scoring protocol for target-specific virtual screening is developed and validated using the challenging Directory of Useful Decoys (DUD) data sets. In the first step our target-specific pose (-scoring) filter developed in Aim 2 is used to filter out/penalize putative pose decoys for every compound. Then in the second step the remaining putative native-like poses are scored with MedusaScore, which is a conventional force-field-based scoring function. This novel screening protocol can consistently improve MedusaScore VS performance, suggesting it possible applications to practical pharmaceutically relevant targets.

Acknowledgements

I would like to sincerely thank the following people:

To Dr. Alexander Tropsha for his scientific guidance, support, and patience

To other members of my committee: Drs. Nikolay Dokholyan, Michael Jarstfer, Stephen Frye, and Scott Singleton for their time and valuable comments on my dissertation

To Drs. Simon X. Wang, Zheng Yang, Alexander Golbraikh, Shuangye Yin, Shubin Liu for their time and effort in assisting and guiding my research projects

To Drs. KH Lee and Weifan Zheng for their warm encouragement and help

To Dr. Alexander Sedykh for his scientific inspiration, care, and patience

To my dearest family for their endless support

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AD	Applicability Domain
ADME	Absorption, Distribution, Metabolism, Excretion
CCR	Correct Classification Rate
CSAR	Community Structure Activity Resources
CV	Cross Validation
DFT	Density Functional Theory
DUD	Directory of Useful Decoys
EI	Efflux Index
EN	Electronegativity
EPI	Efflux Pump Inhibitor
FEP	Free Energy Perturbation
HTS	High Throughput Screening
LIBSVM	Library for Support Vector Machines
LIE	Linear Interaction Energy
LMO	Leave Multiple Out
LOO	Leave One Out
MCT	Maximal Charge Transfer
MDR	Multiple Drug Resistance
MIC	Minimum Inhibitory Concentration
MLR	Multiple Linear Regression
NCE	New Chemical Entity
NMR	Nuclear Magnetic Resonance

kNN	k-Nearest Neighbors
PDB	Protein Data Bank
PLS	Partial Linear Squares
QSAR	Quantitative Structure Activity Relationship
QSBAR	Quantitative Structure Binding Activity Relationship
QSPR	Quantitative Structure Property Relationship
RMSD	Root Mean Square Deviation
SAR	Structure Binding Activity Relationship
SBDD	Structure-based Drug Design/Discovery
SVM	Support Vector Machines
TC	Tanimoto Coefficient
VS	Virtual Screening

Chapter 1 Introduction:

Drug discovery is an expensive and time-consuming process, starting from target protein establishment to FDA approval. According to the statistics in pharmaceutical industry during the 1990s, it took an average of 14 years and cost around $800 million to bring a new drug into the market.[1] In addition to inevitably lengthy and expensive clinical phases where roughly nine in ten compounds is forced to discontinue,[2] the attrition rate in discovery stage is also similarly high. This demands more efficient methods to identify new chemical entities (NCE) to maintain a profitable pharmaceutical company. Mostly, the high-throughput screening (HTS) approach is applied to identify initial chemical hits from a large compound collection in the early stage of drug discovery process.[3] However, HTS approach is also a costly campaign, involving automated robotic screening systems, large quantities of resources and time-consuming assay set-ups, which only large pharmaceutical companies or few highly specialized labs can afford[4]. Thus, to complement the experimental HTS approach in the hope of speeding up the discovery rate in hit-identification phase, computational methods, for example, virtual screening (i.e., searching libraries *in silic*o and selecting only a limited number of molecules for testing), are suggested and applied in order to identify hits with higher reliability yet less effort.[5-7] Moreover, computational methods can also be applied in the lead-optimization phase,[8] where dozens of compounds need to be synthesized/tested to achieve desired ADME-Tox properties and sufficient binding affinity for the target protein. Modeling based on a set of tested compounds can help predict the potency of unknown compounds in order to prioritize the synthesis. To date, numerous

computational methods have been extensively applied in the drug discovery process with varying degree of success. This chapter will provide an overview of these technologies and some well-known limitations as well as the outline of this dissertation.

1.1 Cheminformatics in Drug Discovery

The term cheminformatics is firstly introduced in literature by Brown.[9] Despite various extensions,[10] broadly speaking, cheminformatics can be defined as: the application of informatics methods to solve chemical problems. Traditionally, the subjects in cheminformatics are mainly associated with small molecules despite the fact that macromolecules such as proteins and DNAs are also considered as chemicals. The research in understanding the relationship between macromolecules and ligands (mostly small molecules) is usually discussed intensively in the realm of structure-based drug design (**Chapter 1.2**).

The scientific roots of cheminformatics in drug discovery[11, 12] can be traced back to the pioneering work conducted by Hansch and Fujita who *quantitatively* explain the property of a series of structurally related small molecules based on their steric, electrostatic, and hydrophobic effects.[13] This is the so-called quantitative structural-activity analysis (QSAR) or quantitative structural-property analysis (QSPR). The basic approach to the issue of predicting properties can be simplified to this equation: P = f(C), where the molecular property P is expressed as the mathematical function of molecular structure C. The implicit assumption of this equation is that compounds with similar structures should have similar property profiles. To date, the general QSAR modeling procedure can be summarized as follows: data preparation, data analysis, and model validation.[14] The resulting models can be used to explain the relationship between the molecular property and the chemical features in

molecules, helping to design new molecules, or identify molecules with desired property through searching chemical databases (i.e., database mining).

Starting from the data preparation, a high-quality data set with reliable property measurements is the prerequisite to constructing predictive models. After the curation of data set,[15] which means the information of small molecules are extracted and converted to electronic formats, small molecules can be further represented by a set of parameters called molecular descriptors. Descriptors can be generally divided into 1D, 2D, or 3D descriptors, depending on the dimensionality of molecular representation where they are calculated. For example, the molecular mass descriptor or count-of-hydrogen-donor descriptor is considered as 1D descriptor; the topological indices descriptors which are widely applied in 2D-QSAR modeling take account of atom connectivity derived from the 2D chemical graphs; the molecular surface descriptors are 3D descriptors, whose values are dependent on the experimental/predicted active 3D conformation. Moreover, if the geometry of target protein (receptor) is available, the enthalpy contributions, which are calculated based on the interactions between protein and small molecules, can be treated as descriptors in receptor-dependent (RD) 3D-QSAR modeling. Zhang *et al.* in our laboratory employs a different approach to incorporating the protein-ligand chemical geometrical knowledge into descriptor calculation. The descriptors are coined as ENTess descriptors,[16] which are obtained by using Pauling electronegativity (EN) as atomic property and Delaunay Tessellation (Tess) to characterize the protein-ligand interface. The ENTess descriptors have been successfully applied in constructing quantitative structure-binding affinity relationship (QSBAR) models for 264 x-ray characterized protein-ligand complexes with known binding affinity. The

extension of ENTess descriptors – PL/MCT-tess descriptors – will be discussed in this dissertation (**Chapter 3**).

After completing the stage of data preparation, the next stage deals with the selection of techniques which optimize the correlation between desired property (dependent variable, Y) and molecular descriptors (independent variables, Xs) during model training. These optimization techniques can be generally divided into *linear* or *non-linear*, depending on whether the equation that is applied to explain the relationship between Y and Xs, is a linear combination of parameters or not. The most extensively applied linear method in QSAR studies is Partial Least Squares (PLS), which extends the traditional multiple linear regression (MLR) method when the number of independent variables (descriptors) is much larger than the number of data instances, a common situation in modern QSAR. However, as the increasing availability of experimental data resources, more and more compounds with diverse scaffolds are incorporated into QSAR modeling, the assumption that the variance of independent variables *linearly* corresponds to the variance of dependent variables is not always true. Instead, non-linear models may be constructed using machine learning algorithms such as *k* nearest neighbors (*k*NN). The *k*NN method is firstly introduced to QSAR world in 2000,[17] where a particular compound's property is predicted by its *k* nearest neighbors defined in a subset of descriptor space (resultants from the variable selection optimization).

Furthermore, sometimes researchers care more about whether unknown compounds have certain property or not (e.g., active or inactive) or if they can classify unknown compounds into groups (e.g., high affinity, medium affinity, or low affinity). By contrast, the classification algorithms are employed to construct binary or multi-class classification QSAR

models. The support vector machine (SVM) and random forest (RF) are among the most popular classification techniques. For example, the SVM algorithm searches for the optimal hyperplane that separates the two classes in the descriptor/feature space by maximizing the distance (called *margin*) between the classes' closest points. If the data is not linear separable in the descriptor space, the *kernel trick* is applied to project the data into higher dimensional feature space where the linear separation may exist.

Applying rigorous model validations after and during the model construction is important and necessary to afford predictive QSAR models.[18, 19] The five-fold external cross validation technique or at least a set of randomly removed compounds preserved *solely* for validation should be conducted. For five-fold external validation, the modeling set is divided, by random selection, into five nearly equal subsets. Each subset will be used solely as external set for the models built from the remainder. On the other hand, during the model construction, as emphasized in the previous study,[20] training-set-only modeling is insufficient to achieve models with validated predictive power when using the leave-one-out (LOO) procedure, which each compound is excluded once for validation while remainders are applied for training. Thus, additional internal test sets are needed to construct predicted QSAR models.[20] Moreover, the Y-randomization validation test should be conducted, where the performance of mock models constructed with randomly shuffled independent Y variable (response) is compared to that of the 'real' models under the same modeling protocol. All of the validated models performed significantly better than the randomized models are eligible for external prediction. When predicting the external set, the consensus prediction technique, which is carried out by averaging the predicted activity values resulting from all eligible models, usually has better prediction accuracy comparing with by using the result from a

single, “best” model. This could be resulted from the prediction error of compounds from one model cancelled by the correct predictions from the other model if the errors do not correlate. Furthermore, the compound should be predicted only when it is similar to the training set molecules (i.e., those within model applicability domain).[18]

Naming by the dimensionality of descriptors applied for model building, two types of QSAR methods, 2D-QSAR and 3D-QSAR are regularly compared to each other in many aspects such as the model performance and the ease of descriptor interpretation. Compared to 2D-QSAR models, 3D-QSAR models are more easily interpretable due to its visualizablity, making it simpler to suggest compounds for synthesis. The most popular commercial 3D-QSAR methods include Catalyst[21] and Phase[22]. However, a recent study comparing these two programs demonstrates that the prediction accuracy of external validation set is less acceptable (the squared correlation coefficient, R^2, less than 0.5), informing the further development of 3D-QSAR methodology is necessary.[23] By contrast, our laboratory has been working on developing predicted 2D-QSAR workflow. Using predictive QSAR models as a virtual screening tool in hit discovery, many success stories are published.[24-26] Herein, the QSAR binary classification modeling approaches are applied in several presented projects such as AmpC β-lactamase and Gram-negative efflux property (**Chapter 2**). And an extension of ENTess descriptors, P/L MCT-tess descriptors, is applied in QSBAR model building and structure-based virtual screening (**Chapter 3** & **Chapter 4**).

1.2 Structure-based Drug Design

Structure-based drug design/discovery (SBDD) is defined as the use of 3D target protein structural information in the development of biologically active molecules (e.g., drugs or drug candidates).[27] By understanding the interactions between the target protein and

molecules, chemists could rationally design and optimize the lead molecules compared with time-consuming systematic modifications of molecular structures by cycles of trial-and-error. The 3D protein structural information can come from X-ray crystallography, NMR spectroscopy, cryo-electron microscopy, and homology modeling, where 3D protein structural model is constructed based on its amino acid sequence and a related homologous protein structure determined by experiments (e.g., x-ray crystallography). The popularity of SBDD has substantially increased in recent years since the first seminal paper was published in 1982 by Kuntz's group.[28] This mainly results from the remarkable technical advances in determining target protein structures and protein/ligand complexes, multiplying structural resources related to therapeutically relevant target proteins. The solved 3D structures may be deposited in Protein Data Bank (PDB)[29], where researchers can freely search and download structures. The exponential increase of deposited PDB structures since 1980s (from 70 to 64,357 as of April 2010) also raises the quality issue of the applied structures in SBDD. Scrutinizing the congruence between the experimental electron density map and the fitted protein model gradually becomes a common and necessary procedure before any further SBDD calculations. Thus, various subdivided libraries of PDB are curated for such purpose. For example, PDBBind database[30, 31] is curated by culling from high-quality protein-ligand complexes with experimentally measured binding affinity data. Nevertheless, SBDD approaches have been becoming indispensable tools in the early stage of drug discovery process.

The SBDD approaches are usually divided into two classes: docking and *de novo* design. In this dissertation, only the respect of docking is discussed. Docking is defined as the prediction of compound conformations and orientations (i.e., pose) within the target

protein binding site. This process involves several steps: a) search algorithms explore the possible binding regions for each compound within the target protein binding site and generate multiple poses; b) scoring functions are applied to calculate a score for each pose, which is represented the degree of complementarity to the binding site, or the predicted binding affinity; c) the best ranking pose is commonly selected to represent the binding of that particular compound (i.e., binding mode).

Since the pioneering docking program DOCK[28, 32] published in 1980s, a series of other programs, such as FlexX,[33] GOLD,[34] and AutoDock[35], have emerged. Each of them varies in the respect of pose generation algorithms, the applied scoring functions, and the degree of protein/ligand flexibility taken into account. At present, all docking programs allow compounds to dock flexibly, either by exploring the translational and orientational degrees of freedom of pre-generated conformers (e.g., Fred[36]), or by generating the poses on-the-fly (e.g., AutoDock). However, explicit protein flexibility (i.e., the movement of protein backbone/side-chain) is still not regarded as a norm in docking considering the size and possible degrees of freedom of macromolecule.

The SBDD approaches have been successfully applied in several cases. There are two prominent drug discovery projects: the structure-based design of neuraminidase inhibitors[37] contributes the birth of first anti-influenza virus drug – Relenza, and the development of HIV protease inhibitors used as anti-AIDS drugs.[37, 38] Moreover, Aggrastat, a fibrinogen receptor (GP IIb/IIIa) antagonist launched in 1998, is cited as first marked drug whose discovery highly influenced by virtual screening.[39] While these success stories accompany the burgeon of docking programs, the SBDD practitioners are eager to know if the state-of-the-art docking programs can really help them when dealing with novel targets. Since 2000, a

plethora of studies comparing the performance of different docking programs have been published.[40-46] The 2006 GSK paper by Warren *et al.*[47] concludes the current achieved status in docking by conducting an extensive retrospective study using ligand/decoy sets with experimentally determined binding affinities against a wide range of pharmaceutically relevant protein targets. In total, they evaluate 10 docking programs and 37 scoring functions and summarize the performance of those docking programs on three tasks: a) search algorithms in docking *can* generate poses which are closed to the experimentally determined binding mode (native pose) yet less successful in predicting the correct binding mode of ligands; b) docking/scoring can identify ligands among a set of pharmaceutically relevant decoys in virtual screening campaigns but the performance is highly target-dependent; c) in terms of lead optimization, none of the docking programs or scoring functions can make a useful prediction of ligand binding affinity. All of these retrospective studies demonstrate that significant improvements are needed for current scoring functions (or scoring schemes in virtual screening).

In general, scoring functions can be classified into three types[27]: a) force-field based scoring functions rely on explicitly computed electrostatic and van der Waals interaction energies (i.e., enthalpic effects) between the ligand and the protein based on a molecular force field. For example, G-score[48] which is based on the Tripos force field[48] and AutoDock 3.05 score based on the AMBER force field[49]; b) empirical scoring functions are defined as the sum of individual uncorrelated energy terms whose coefficients are optimized from regression analysis by fitting the experimental data such as binding energies/affinities. Several non-enthalpic contributions can be included such as deformation and hydrophobic terms in XSCORE[50] and the ligand rotor term in ChemScore;[51] c) knowledge-based scoring

functions are designed based on various statistical parameters derived from x-ray crystal structures that could reflect the interactions between a ligand and its receptor depending on their molecular environment. Simple distance-dependent pair potentials and non-polar surface-dependent singlet-potential are used in DrugScore.[52] They could implicitly capture the binding effects that are difficult to model in force field based scoring functions. Moreover, consensus scoring schemes[53-56], which various data fusion approaches are used to combine information from multiple scoring results in the hope of compensating the errors inherent in each single score, are also widely employed. However, a paper published in 2005 demonstrates that consensus only works when each of the individual scoring functions has relatively high performance and the scoring characteristics of each individual scoring function are quite different.[56]

A popular strategy to improve the force-field scoring functions is to consider the entropic and solvation effects that are ignored in most of the current force-field scoring functions. The well-known methods are: MM-GB/SA,[57] Linear Interaction Energy (LIE),[58] and Free Energy Perturbation (FEP).[59] However, due to computationally intensive calculations, the application of these methods is constrained for target-specific lead optimization. Moreover, a recent study demonstrates that the prediction accuracy of binding affinity of static x-ray structures is less acceptable[60], no wonder the same poor prediction accuracy is observed when conducting cross-docking calculations where protein induced-fit effects are encountered. The result suggests that, for other than some computationally intensive approaches recently being developed for target-specific lead optimization in structure-based drug design,[58, 59] the improvement of current scoring functions for generic high-throughput molecular docking is also needed.

Recently, a hybrid (empirical + knowledge-based) scoring function incorporating the cheminformatics concepts into conventional structure-based scoring functions is developed in our laboratory.[16] The scoring function is quantitative structure-binding affinity relationship (QSBAR) models constructed by 264 x-ray protein-ligand complexes with known binding affinity using ENTess descriptors. The ENTess descriptors are generated based on the tetrahedra resulting from Delaunay tessellation (Tess), characterizing the protein-ligand interface by means of Pauling electronegativity (EN) values. The output of ENTess scoring function can be directly related to absolute binding affinities and could implicitly take into account binding effects that are difficult to specify, combining the merit of both empirical and knowledge-based scoring function. However, the performance of ENTess scoring function in practical virtual screening is limited. One of the possible reasons could be the limitation in applicability domain of ENTess models. In **Chapter 3**, we report the study managing to improve the ENTess scoring function. .

Regarding the issue of virtual screening in hit identification stage that multi-purpose scoring functions cannot perform consistently across diverse targets, a popular alternatives is to address the problem by including knowledge of a single specific protein target or a family of targets during docking/scoring. For instance, some studies have demonstrated that target-specific customized scoring functions[61, 62] are effective methods for improving the discrimination between true ligands and binding decoys in VS for the aimed target. On the other hand, because the awareness of the scoring problem may originate current scoring functions are solely optimized by existing ligand data, scoring function developers manage to include decoy compounds (or poses) during scoring function optimization.[63-69] In **Chapter 4**, we report the target-specific cheminformatics-based pose (-scoring) filter trained to

discriminate native-like poses of ligands *vs*. pose decoys. The pose filter, along with MedusaScore (a force field-based scoring function), is combined to develop a novel two-step protocol for target-specific virtual screening with the aim to improve the hit enrichment in structure-based virtual screening.

1.3 Summary

This dissertation will describe the contributions to the field of SBDD by incorporating cheminformatics concepts into structure-based scoring methods. Firstly, cheminformatics practices are applied to those problematic cases which conventional structure-based scoring functions find difficult (anti-bacterial leads efflux study) or even fail to address (AmpC β-lactamase study). Secondly, novel *pose* and *binding* structure-based scoring functions and a two-step scoring protocol are developed by employing cheminformatics approaches to improve protein-ligand binding affinity prediction and structure-based virtual screening respectively.

Chapter 2 discusses two case studies demonstrating that cheminformatics approaches can complement structure-based drug discovery/drug design and identify promising hits by virtually screening molecular libraries.

The first case study is prediction of efflux properties (low *vs*. high) for Gram-negative bacteria, by the binary classification QSAR approach with pharmacophore fingerprint descriptors. Bacterial efflux properties are difficult to model by structure-based methods due to the structural complexity of the efflux pump. However, I successfully construct QSAR models which show high prediction accuracy in both internal and external validation. After applying the models to virtual screening, many compounds predicted as low-efflux were experimentally confirmed as such.

The second case study is differentiation of AmpC β-lactamase binders *vs.* binding decoys using the binary classification QSAR approach. The binding decoys are false positives mispredicted by conventional structure-based scoring functions. To differentiate them, I successfully construct predictive QSAR models based on rigorous internal and external validations. Applying the models to predict false positives and false negatives from high throughput screening, the models discard false positives and can rescue false negatives.

Chapter 3 explains the development of *binding* scoring function, which is a collection of QSBAR models. Compared with the previous ENTess scoring function, the new binding scoring function is constructed with an increased number of protein-ligand complexes and novel protein-ligand interfacial descriptors incorporating conceptual DFT atomic properties. Upon the application of global applicability domain, this new binding scoring function shows acceptable prediction accuracy (the squared correlation coefficient: 0.57) towards the Community Structure-Activity Resources (CSAR) data set.

Chapter 4 describes the development of the target-specific *pose* (-scoring) filter with the aim to improve the hit enrichment in structure based virtual screening. The pose filter is developed for each target by building binary classification models that can discriminate native-like poses of ligands *vs.* pose decoys. Furthermore, a two-step scoring protocol for target-specific virtual screening is developed. In the first step our pose filter is used to filter out/penalize putative pose decoys for every compound, and in the second step the remaining putative native-like poses are scored with MedusaScore, which is a conventional force-field-based scoring function.

Chapter 5 presents conclusions and future studies.

Chapter 2 Cheminformatics Approaches Complement Structure-based Virtual Screening:

2.1a Classification of Gram Negative Bacteria Efflux Properties of Antibacterial Leads Using Pharmacophore Fingerprint-based SVM QSAR Modeling and Application to Virtual Screening

2.1a.1 Introduction

Bacterial multidrug resistance (MDR) is frequently reported in clinics, underlining the need for developing new antibiotics. Unfortunately, a majority of recently approved antibiotics and developing compounds still lack activities against Gram-negative bacteria despite the fact that they cover a number of novel, well-conserved bacterial targets to overcome the resistance by target modification and enzymatic drug inactivation. It is widely recognized that this intrinsic resistance largely results from the constitutive expression of multi-drug efflux pumps. Unlike other bacterial efflux pumps only selectively extruding specific drugs, efflux pumps involved in MDR can pump out a number of antibiotics with diverse structures and unrelated functions, rendering simultaneous bacterial resistance. The efflux issue is especially serious for Gram-negative bacteria due to combined effects of efflux pumps to actively expelling antibiotics (efflux) and the unique Gram-negative bacteria's outer membrane to reducing antibiotics uptake (influx)[70] (**Figure 2.1a.1**), providing an effective barrier against both hydrophilic and hydrophobic antibiotics.[71]

There are a total of five families of efflux pumps associated with MDR:[72] the ATP binding cassette (ABC) superfamily, the major facilitator superfamily (MF), the multidrug and toxic-compound extrusion (MATE) family, the small multidrug resistance (SMR) family, and the

resistance nodulation division (RND) family. Among them, members in the RND family are the most significant efflux determinants of intrinsic and acquired resistance in Gram-negative bacteria. Efflux pumps in RND family are organized into three-component structures transversing both inner and outer membranes, allowing diverse antibiotic substrates to directly expel out from the cytoplasm and periplasm (**Figure2.1a.1**). Recent x-ray structures of efflux pumps co-crystallized with several ligands simultaneously in an extremely large cavity confirms the diverse substrate specificity of efflux pumps.[73]

The inhibition of efflux pumps has been suggested as a viable approach to overcome MDR. Common strategies for efflux inhibition include a) development of efflux pump inhibitors (EPI) in combination with available antibiotics to increase antibacterial potency; b) design of anti-bacterial lead compounds which can elude efflux pumps. The latter approach is potentially more practical compared to the former one, where EPIs can only be effective when the complimentary antibiotic substrates share the same binding site (i.e., competitive inhibition). One of the goals of this study is to help medicinal chemists to identify anti-bacterial lead compounds that can elude the efflux pumps using *in silico* models.

Computer-aided drug design, including both structure-based and ligand-based drug design, has demonstrated its contribution in drug discovery.[37] Despite the availability of x-ray crystal structures of each component of *E. coli* efflux pump,[74-76] the complexity of efflux pump stoichiometry and the extra-large substrate binding site make structure-based approach less plausible. On the other hand, modeling of efflux properties is also challenging using ligand-based approach since efflux is a dynamic process of compounds going through cellular trans-membrane channels instead of binding to a static target. Published ligand-based studies mostly fall into the realm of explanatory models with fairly small data sets (<45

compounds).[77, 78] Herein, *in silico* quantitative structure-activity relationship (QSAR) models are built to classify Gram-negative bacteria efflux properties (low *vs.* high) with five-fold external cross validation (CV) average accuracy as high as 79%. The predictive models are built and validated using GlaxoSmithKline (GSK) in-house pharamacophore fingerprint (pFP) descriptors and support vector machine (SVM) algorithm on a GSK proprietary *K. pneumoniae* efflux data set (~400 compounds). The models are subsequently applied in virtual high throughput screening and 60 out of 75 available potent computational hits are confirmed as low-efflux by bioassays, achieving high accuracy of 80%. The predictive models can also further be used to prioritize synthesis of GSK anti-bacterial series. These encouraging results provide a good template for further efflux modeling research.

2.1a.2 Methods

2.1a.2.1 Data Sets

The GSK anti-bacterial series are broad-spectrum bacterial topoisomerase IIa inhibitors **(Figure 2.1a.2)**, which show a novel mode of inhibition different from clinical topoisomerase IIa inhibitors in the quinolone series, providing the hope of against topoisomerase IIa - mediated cross-resistance.[79] However, the novel GSK anti-bacterial series also suffer from the multidrug efflux pump issue especially in Gram-negative bacteria.[80] Thus, it is desirable to build *in silico* models to predict efflux properties of GSK anti-bacterial series to further improve potencies by reduction of efflux activities.

Minimum inhibitory concentration (MIC) assays are used to measure the *in vivo* antibacterial activity (μg/ml) of compounds. The efflux properties of anti-bacterial compounds are quantitatively characterized by the efflux index (EI), which is defined by the

ratio of the MIC of wide-type bacteria to that of efflux knock out bacteria. The EI values are transformed to the logarithmic value (pEI) for modeling purpose (**Equation 2.1**).

$$p(\textit{Efflux Index}) = \log_2(\frac{\mathrm{MIC}_{\mathrm{WT}}}{\mathrm{MIC}_{\mathrm{KO}}}) \quad (2.1)$$

Each MIC is measured at least twice to ensure data integrity. The experimental variability of pEI for each compound is usually around ±1, but could be as high as ±2. Therefore, a classification model is better suited for modeling the efflux indices. All compounds are classified by a threshold of pEC = 6, i.e. 64-fold difference between wide type MIC and efflux knock-out MIC, determined by biological interests. In total, there are 399 GSK anti-bacterial series annotated as low and high efflux properties against *K. pneumoniae* for modeling, containing 149 low-efflux compounds (37%) and 250 high-efflux compounds (63%). The pEI distribution of dataset is shown in **Figure 2.1a.3**. Compound structures are relatively diverse, including, for example, the tricyclic nitrogen series[81] and the cyclohexane/cyclohexene series.[82]

Furthermore, seventeen compounds in a new subseries outside of training set are served as an additional external validation set. The pEI distribution of these 17 compounds is shown in **Figure 2.1a.4**.

Regarding the library used for virtual screening, a total of 4013 historical GSK bacterial topoisomerase IIa inhibitors without *K. pneumoniae* efflux properties are curated to search for low-efflux templates.

2.1a.2.2 Training, Test, and External Validation Set Selection

The overall modeling workflow is shown in **Figure 2.1a.5**. The double five-fold cross validations are applied for modeling, including internal (optimization of model parameters) and external cross validations (testing model performance). The data set is randomly split into five subsets, where the ratio of low-efflux to high-efflux compounds corresponding to that in the modeling set. Each subset is applied independently to validate the models built from the remaining subsets by five-fold internal CV using the Support Vector Machines (SVM) algorithm. Furthermore, seventeen compounds in a new subseries outside of training set are served as an external validation set. The pEI distribution of these 17 compounds is shown in **Figure 2.1a.4**.

2.1a.2.3 Generation of Pharmacophore Fingerprint (pFP) Descriptors and pFP Noise Reduction by pFPBitRank Tool

The GSK in-house pFP program is applied to generate the pFPs for all compounds. The implementation is similar to the one previously reported.[83, 84] There are six pharmacophore feature types (hydrogen bond donor/acceptor, positive/negative ionizable, hydrophobic centroid, and aromatic centroid) and seven inter-pharmacophore distance bins (1.0-3.0, 3.0-4.0, 4.0-5.2, 5.2-6.5, 6.5-8.0, 8.0-10.0, 10.0-50.0, unit: Å). The distance bin boundaries are statistically determined to produce equal occupancies across a large set of GSK drug-like compounds. The combination of six pharmacophore feature types and seven distance bins leads to a total of 19208 pFP bits.

For each compound, multiple conformers are generated by using Omega [36] (version 2.2) with default parameters. Pharmacophore features are assigned to atoms and functional groups in each conformer, while inter-pharmacophore distances are classified into seven bins. Then all three-point phamacophore triangles are enumerated based on the triangle constraint

and stored into a bit string to generate per-conformer pFP (**Figure 2.1a.6**). All per-conformer pFPs of one compound at each pFP bit position are further processed by logical OR to generate the per-molecule pFP of that compound. As long as a certain three-point pharamacophore appears in any conformer, that bit is turned on in the per-molecule pFP. These three-point pharmacophore fingerprints are applied as descriptors to capture the physiochemical nature of a compound, as the GSK pFP descriptors have been successfully applied in both lead optimization and lead identification projects.[85-87]

The pFP BitRank tool [85] is applied to increase the signal to noise ratio in pFPs based on information from known low-efflux/high-efflux compounds. A score for each bit is calculated using **Equation 2.2**. Given the pFPs of sets of active (e.g., high-efflux) and inactive (e.g., low-efflux) compounds, each bit is scored according to relative prevalence amongst actives (a) or inactives (i)

$$\text{BitScore}(j) \equiv f^{a}_{1} * f^{i}_{0} + f^{a}_{0} * f^{i}_{1} \qquad (2.2)$$

where j is the id of each bit and f^{x}_{y} is the fraction of bits whose value is y (e.g., 0 or 1) amongst compounds in class x (e.g., active or inactive).

To estimate the "noise" bitscore value, the activity labels are shuffled (y-randomization) and the score for each bit is recalculated. This procedure is repeated 100 times and the mean as well as standard deviation are calculated using all bitscore values from randomization. Only the bits with bitscore higher than a certain Z cutoff are retained and applied as descriptors in SVM model building.

$$Bitscore = \bar{y} + Z\sigma$$

Here, $\bar{y}$ is the average bitscore values from randomization, σ is the standard deviation of these bitscore values, and Z is an arbitrary parameter to control the significance level. (Further implementation details are not disclosed by GSK.)

2.1a.2.4 Support Vector Machine Classification Method

The Support Vector Machines (SVM) algorithm implemented in the open-source LibSVM[88] package are employed to build binary classification models. The SVM algorithm searches for the optimal hyperplane separating the two classes in the descriptor space by maximizing the margin between the closest points of the two classes (**Equation 2.3**),

$$\min \tfrac{1}{2} W^T W + C\sum_{i=1}^{l} \xi_i \text{ , subject to } y_i(W^T \Phi(X_i + b)) \geq 1 - \xi_i \qquad (2.3)$$

where *C* is the penalty parameter and $\xi_i \geq 0$ is the slack parameter. To make the data set linearly separable, the data points are projected to a higher dimensional space by Radial Basis Kernel (RBF),

$$K(x, x_j) \equiv \Phi(X_i)^T \Phi(X_i) = \exp(-\gamma \left\| x_i - x_j \right\|), \gamma > 0 \qquad (2.4)$$

where γ is the kernel parameter. We employ the Python script (grid.py) provided by LibSVM to optimize parameters *C* and γ during model building with 5-fold cross validation (CV). The search range of *C* and γ are -5 to 15 and -15 to 0 respectively.

2.1a.2.5 Virtual Screening Using pFP-SVM Models

All SVM models with eligible CV accuracy are used to predict the test set. When applied to the compounds in the test set of each fold CV, compounds are considered as low (or high) efflux only when they are predicted as low (or high) efflux consistently by no less

than 50% of all eligible models. However, in the VS study, a higher threshold (90%) is applied to select low-efflux compounds for experimental testing. A higher threshold is assumed to have higher confidence in prediction.

2.1a.3 Results and Discussions

2.1a.3.1 Relationship between Distribution Coefficient and Efflux Index

The correlation between the measured logD values (the logarithmic value of distribution coefficient) and the pEI values is analyzed based on the hypothesis that hydrophobic (i.e., high distribution coefficient) compounds tend to have higher EI values due to favorable interactions between hydrophobic compounds and the aromatic binding site of efflux pumps. As shown in **Figure 2.1a.7**, there is a marginal correlation between these two factors, indicating that efflux modeling is more complicated than simple polarity modeling. Therefore, extra structural information (e.g., pFP) is needed to build *in silico* efflux models.

2.1a.3.2 SVM Binary Classification Models

The pFP descriptors are calculated for all compounds and the pFP BitRank tool is applied (Z-cutoff=0.5) to the efflux modeling data set to reduce the "noise" bits according the protocol described in **Method 2.1a.2.3.** In total, there are 2248 pFP descriptors used in model building. Firstly, pFP regression models using partial least squares (PLS) algorithm are constructed to correlate pFP bits with the pEI values. However, only models with poor prediction accuracy are obtained (data not shown). The poor prediction accuracy might result from the large experimental variability (±2) of pEI. Therefore, the binary classification models are constructed instead to differentiate low-efflux compounds from high-efflux ones using SVM algorithm. Only models with internal CV accuracy higher than 75% are saved for

consensus prediction on the test sets. The overall statistics are summarized in **Table 2.1a.1**. The average prediction accuracy for training sets and test sets is as high as 76% and 79% respectively. The consensus models from the 1st fold and the 4th fold are applied to the additional external validation set and in the virtual screening.

2.1a.3.3 External Validations

Seventeen compounds in a new subseries outside of training set are served as an additional validation set for predictive models. Compounds are classified as low (or high) efflux only when they are predicted as low (or high) efflux consistently by no less than 50% of all eligible models. The prediction results are tabulated in **Table 2.1a.2**. Almost all low-efflux compounds (8 out of 12) are predicted correctly and three out of four false positives have pEI equal to 6 (i.e., borderline compounds).

In summary, validation results show that the pFP-SVM models can differentiate pharmacophore features of low-efflux compounds from high-efflux compounds and can be applied for virtual screening.

2.1a.3.4 Virtual Screening Using Predictive pFP-SVM Models

The validated consensus prediction models are employed to virtually screen 4013 historical GSK bacterial topoisomerase IIa inhibitors to identify low-efflux chemical templates against *K. pneumoniae*. A higher consensus threshold (90%) is applied to select low-efflux compounds for experimental testing. In total, there are 280 selected compounds, each of which is predicted as low-efflux by no less than 90% of eligible models. Out of 280 compounds, 115 available compounds are experimentally tested. 35% of them are identified as low potent against *K. Pneumoniae* and therefore their efflux properties cannot be

determined experimentally. For the remaining compounds (75 compounds), 80% are confirmed with low efflux properties.

2.1a.3.5 Virtual Screening Using LogD value

Further efforts are spent to investigate the prediction accuracy of using logD value alone to fish out the low-efflux compounds from those 75 compounds which could be assumed randomly selected from the library. The range of measured logD values of these 75 potent compounds is from -0.3 to 1.5. The probability of fishing out low-efflux compounds based on logD values lying in that range in the modeling set is 0.47 in comparison with the prediction accuracy (0.80) by using pFP-SVM models, demonstrating the benefits of QSAR modeling of efflux properties.

2.1a.4 Conclusions

The pFP-SVM-based efflux classification models are able to differentiate the low-efflux compounds from high-efflux compounds in the GSK anti-bacterial series and can identify the low-efflux compounds in new subseries outside of training set. Over 4000 historical GSK bacterial topoisomerase inhibitors are screened by the models and subsequent experimental validation demonstrates that the models afford high prediction accuracy (80%) of identifying low-efflux structures. It suggests that the models could be a plausible tool for lead optimization and virtual screening. The encouraging results of pFP-SVM-based efflux classification model building, validation, and virtual screening provides a good template for further efflux modeling exercises as well as modeling against endpoints with large experimental variability.

Figures for Chapter 2.1a

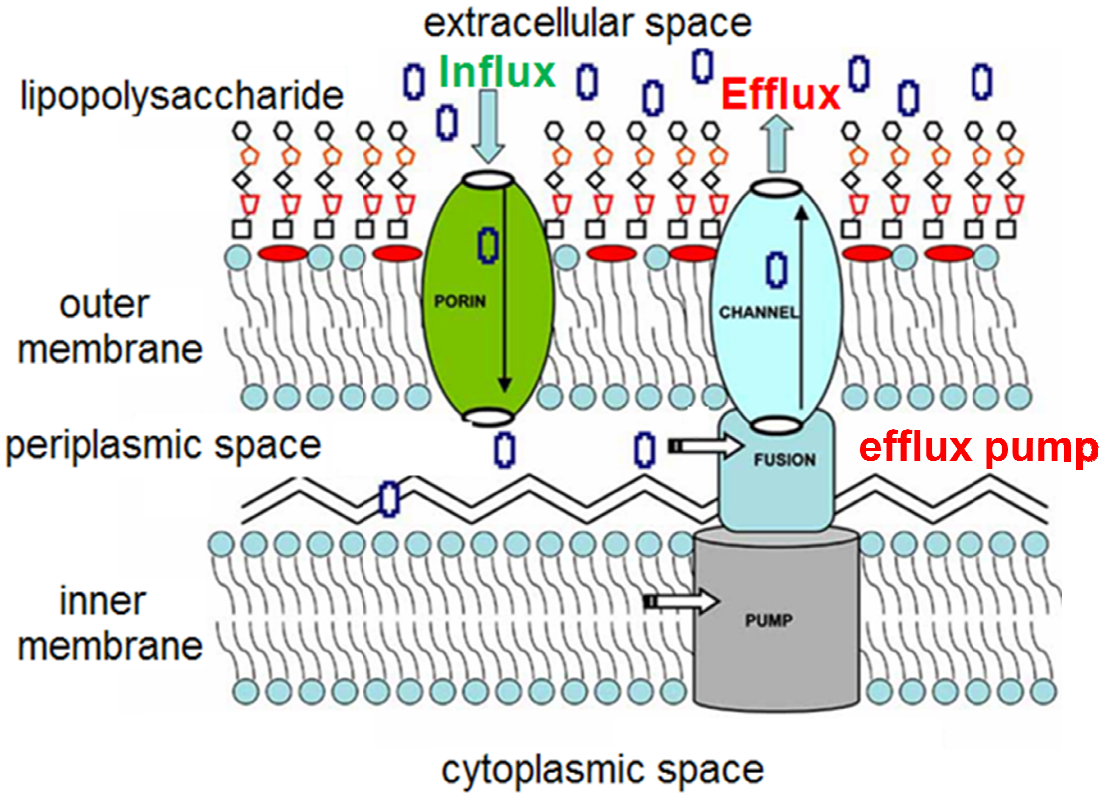


Figure 2.1a.1: Schematic representation of both the inner and outer membrane of Gram-negative bacteria together with the porous layer of peptidoglycan, the main target of beta-lactam antibiotics (in blue).

Influx and efflux systems are also inserted to show the uptake and suggested extrusion of antibiotics. The figure is modified from the Figure (1) in *Curr Drug Targets*. 2008, 9, 779.[89]

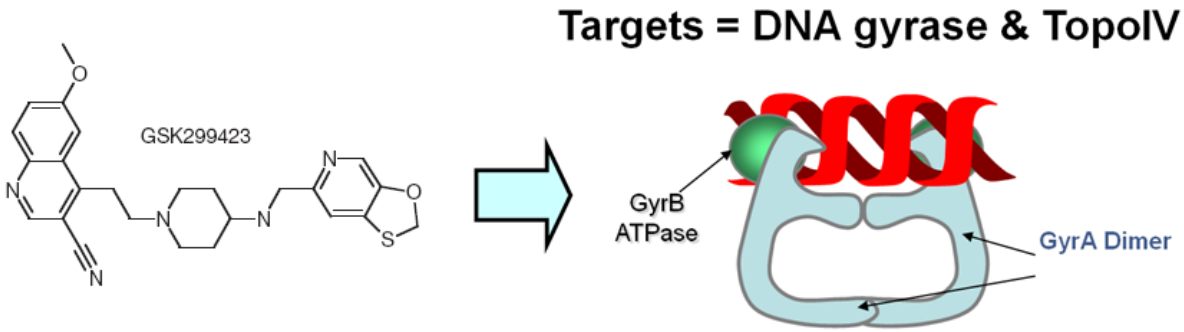


Figure 2.1a.2: The novel GSK antibiotic series target bacterial topoisomerase IIA (DNA gyrase and topo IV).

The inhibition mechanism of GSK bacterial topoisomerase IIA inhibitors is different from the one of fluoroquinolones, circumventing the topoisomerase IIa - mediated cross-resistance.

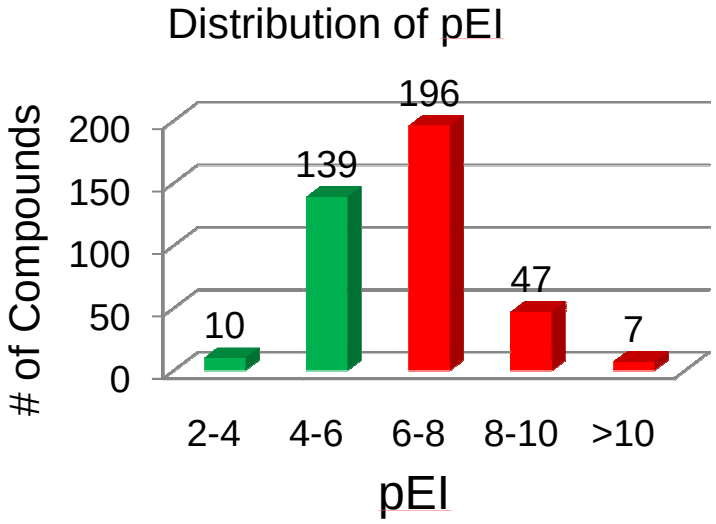


Figure 2.1a.3: The distribution of pEI values of the GSK bacterial topoisomerase IIa dataset.

Applying pEI, 6, as the threshold, totally, there are 149 low-efflux compounds (Green) and 250 high-efflux compounds (Red).

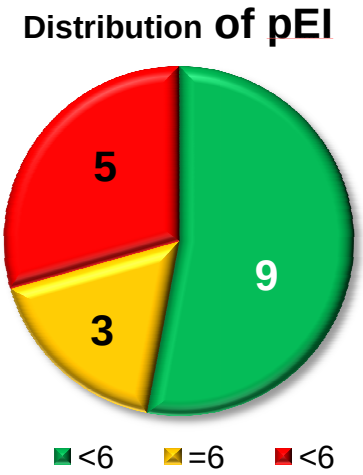


Figure 2.1a.4: The distribution of pEI of the newly synthesized compounds by exploring structure activity relationship (SAR) with single functional group substitution of GSK antibiotics series.

There are nine low-efflux compounds (Green), three borderline compounds (yellow), and five high-efflux compounds.

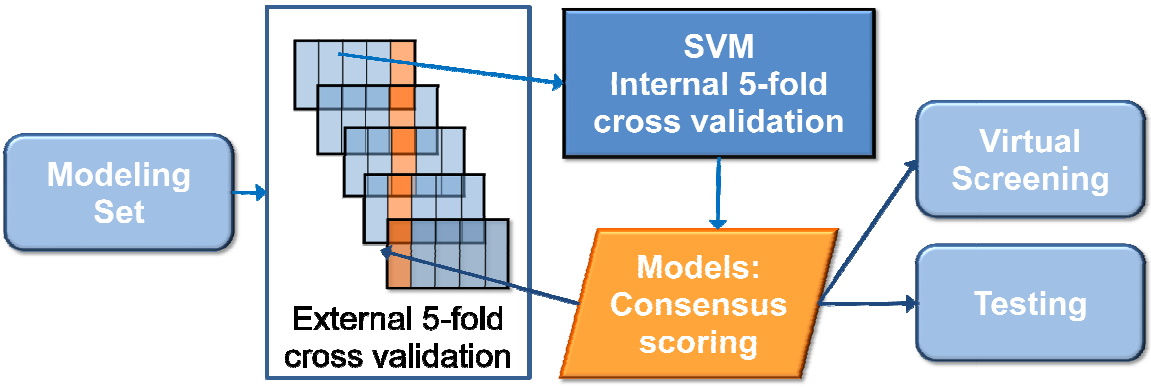


Figure 2.1a.5: The workflow of efflux model building, validation, and virtual screening.

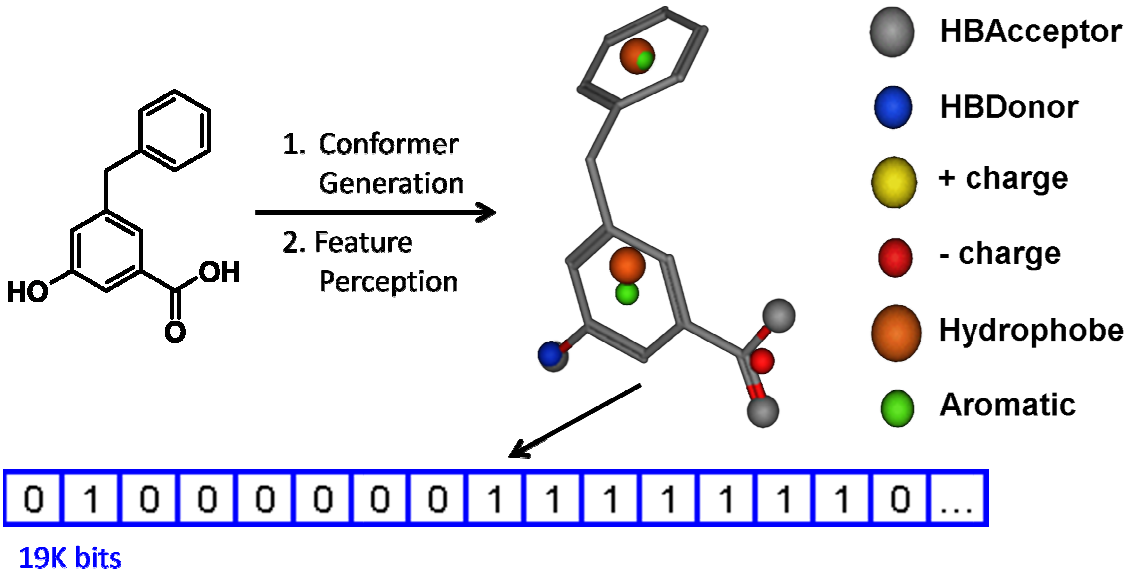


Figure 2.1a.6: The pFP descriptor calculation. For each compound, multiple conformers are generated, each of which is assigned six pharmacophore features.

Along with seven distance bins (1.0-3.0, 3.0-4.0, 4.0-5.2, 5.2-6.5, 6.5-8.0, 8.0-10.0, 10.0-50.0, unit: Å), the 19208 three-point phamacophore features are enumerated based on the triangle constraint and stored into a bit string.

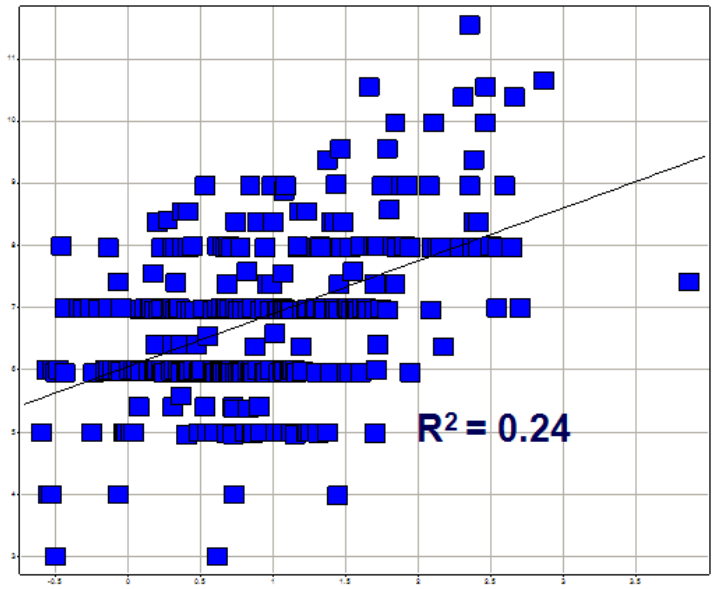


Figure 2.1a.7: The correlation plot of pEI values (x-axis) and logD values (y-axis). There is only marginal correlation between pEI values and logD values ($R^2 = 0.24$).

Tables for Chapter 2.1a

Table 2.1a.1: The statistics of accuracies from internal and external five-fold cross-validation (CV) by models with internal CV accuracy larger than 75%.

CV split	Average CV accuracy (%)	Test set accuracy (%)	Low efflux accuracy (%)	High efflux accuracy (%)	# of models w/ accuracy ≥ 75%
#1	75	84	71	91	50
#2	75	80	76	82	17
#3	76	78	71	84	20
#4	78	76	65	82	100
#5	75	77	63	82	6

Table 2.1a.2: The confusion matrix of 17 newly synthesized compounds with single functional group substitution.

predicted / experimental	Low (pEI <=6)	High (pEI > 6)
Low (pEI <6)	8	1
Borderline (pEI = 6)	0	3
High (pEI > 6)	0	5

2.1b Differentiation of AmpC β-Lactamase Binders vs. Binding Decoys Using Classification *k*-NN QSAR Modeling and Application of QSAR Classifier to Virtual Screening

2.1b.1 Introduction

Due to rapid advances in protein crystallography[90, 91], the number of x-ray characterized biological targets and their complexes with various low molecular weight ligands in the RCSB Protein Data Bank (PDB)[29] has been growing rapidly. This growth has been concurrent with the development of a vast array of structure-based virtual screening approaches [92-101]. These methods include two critical components, i.e., docking and scoring. It has been shown that multiple binding poses of putative receptor ligands resulting from docking include those that are geometrically close to the native (i.e., experimental) ligand orientation in the binding site. However, identifying (the most) native-like binding poses among many alternatives (i.e., 'geometrical decoys') resulting from docking continues to present a universal problem to most scoring functions.[47, 102, 103] Furthering this problem is a demonstrated inability of many scoring functions to discriminate between ligands that are known to bind to the target receptor from those known to be non-binders yet predicted to bind by a docking/scoring method (so called 'binding decoys') [42, 104].

Many strategies have been proposed to improve scoring functions such as to decrease the number of false positives as well as improve the enrichment of true positives[42, 105-109]. Nevertheless, in a recent study, Shoichet and coworkers[103, 110] reported docking of over 200,000 compounds into the active site of AmpC beta-lactamase that produced many binding decoys. These compounds were ranked highly by many scoring functions such as DOCK, ScreenScore and FlexX etc., but were found to be false positives as a result of experimental

validation. Similar results have been observed for several other systems (available from the B. Shoichet's laboratory website, http://shoichetlab.compbio.ucsf.edu/take-away.php).

From the traditional three-dimensional docking and scoring prospective, the existence of binding decoys illustrates the need for developing more robust scoring functions. However, the same results could be also approached from a cheminformatics prospective. Thus, the two groups, i.e., experimentally confirmed binders and binding decoys represent two classes of compounds that could be possibly discriminated by their chemical features, or descriptors. The problems of this type (i.e., discriminating binding from non-binding compounds based on their chemical structure descriptors only) are rather common in case of ligand based drug design approaches such as Quantitative Structure Activity Relationship (QSAR) modeling. In fact, the use of binary QSAR modeling towards the problem of discriminating true binders vs. decoys may be perhaps even more challenging than the standard binary QSAR modeling. Indeed, in this case both classes of compounds are apparently sufficiently similar to each other to fit into the same receptor binding site to be indistinguishable by well-defined and validated scoring functions. Thus, being able to discriminate binders vs. (similar) decoys should be a difficult exercise. On the other hand the successful structural models could potentially inform protein structure based scoring functions about specific functional groups that are primarily responsible for the discriminatory power of the QSAR models but most likely are not adequately scored by the traditional scoring functions. Furthermore structural models could be potentially used for mining external compound libraries to identify novel putative binders providing a potential alternative to structure based virtual screening methods. [12, 111]

The goal of this study was to develop robust binary classification QSAR models that would have high predictive power to differentiating binders vs. non-binding 'decoys' for AmpC beta-lactamase. We have employed a rigorous validated QSAR modeling workflow that has been developed in our laboratory in recent years. This workflow that incorporates a virtual screening module was applied successfully to several ligand datasets leading to the identification of experimentally confirmed novel hits for different biological targets [112-116] (see recent review [117]). Herein, we report on classification QSAR models that are capable of discriminating binders from decoys with the external classification accuracy exceeding 90%. Furthermore, we have used these models to screen the compound library tested earlier in the AmpC assay and available from PubChem [118]. We have identified 15 molecules as putative AmpC ligands and demonstrated in subsequent experimental studies that five compounds chosen from these hits were millimolar binders. It worth emphasizing that in all studies reported in this paper we did not use any information on the crystallographic structure of AmpC-ligand complexes and moreover, chemical descriptors were generated from two-dimensional rendering of molecular structures.

2.1b.2 Methods

2.1b.2.1 Data Sets

Compounds used for QSAR model building.

The AmpC beta-lactamase inhibitors and binding decoys were downloaded from Dr. Brian Shoichet's laboratory web site [119]. This dataset contains 21 confirmed inhibitors (cf. **Appendix I**) and 80 decoys. The inhibitors were shown to be non-covalent, reversible AmpC beta-lactamase inhibitors.[103, 120, 121] All decoys were shown to have no binding to AmpC at 1mM concentration but falsely predicted to bind by multiple scoring functions [120, 122].

Library used for virtual screening.

We used the dataset of 69653 compounds that was screened in the HTS assays for AmpC beta-lactamase inhibition by the National Center for Chemical Genomics (NCGC). The screening results are reported in PubChem as Bioassays AID584 [123] and AID585 [124]. The experimental protocols are described in [125] as well as in the PubChem database. AID584 and AID585 were designed for screening of specific and promiscuous AmpC beta-lactamase inhibitors, respectively. Compounds are classified as having full titration curves, partial modulation, partial curve (weaker actives), single point activity (at highest concentration only), or inactive. Compounds that showed activity in both AID584 and AID585 assays were considered 'true' positives. However, if compounds were only found active in AID585 but inactive in AID584, they were categorized as 'aggregators'. Thus, 64 compounds were identified as 'true inhibitors of' the AmpC beta-lactamase that could be used to test the ability of QSAR model based virtual screening to recover known hits.

2.1b.2.2 AmpC β-lactamase Competitive Inhibitor Assay

The details of enzymatic assays to measure the efficiency of AmpC beta-Lactamase inhibitors were described in detail elsewhere (26, 40). Briefly, the change in initial rate of substrate hydrolysis at increasing concentrations of the inhibitor was monitored and the IC_{50} was obtained using the resulting dose-response curve. The inhibition constant, K_i, was derived from the IC_{50} value using the Cheng-Prusoff equation.

2.1b.2.3 Training, Test, and External Validation Set Selection

We have followed the rigorous QSAR workflow for model building, validation and database mining (**Figure 2.1b.1**) established in our laboratory (see [117] for recent overview).

For classification QSAR modeling, it would be ideal to have the balanced ratio between different compound classes in the modeling dataset. However, the AmpC beta-lactamase binding dataset included 21 inhibitors and 80 decoys, i.e., it is imbalanced with the inhibitors to non-binders ratio of 1:4. In the absence of special statistical treatment, such ratio would skew the prediction accuracy of the classification models. Thus, the distance matrix was calculated in the multidimensional descriptor space for all 101 compounds and similarity search was carried out using 21 inhibitors as queries against the remaining 80 non-binders. 30 compounds were selected from the original 80 non-binders as most similar to 21 inhibitors using Euclidean distance as similarity metric (we note that this treatment makes the task of building the discriminatory binary QSAR models even more challenging. Consequently, these 30 non-binders combined with 21 true inhibitors formed a new balanced dataset for QSAR model building. The remaining 50 "dissimilar" non-binders were retained as an external validation set. Furthermore, 10 compounds (five binders and five decoys) were randomly excluded from the balanced dataset of 51 compounds and formed a second external validation set. The remaining 41 compounds were considered a modeling dataset that was divided into multiple diverse and representative training and test sets using the Sphere Exclusion approach developed in our laboratory earlier [20, 126].

2.1b.2.4 Generation of 2D Molecular Descriptors

The SMILES [127] strings of each compound in AmpC beta-lactamase dataset were converted to 2D chemical structures using the Unity module of the SYBYL software package [128]. The MolConnZ [129] software (version 4.09) was used to calculate a wide range of topological indices of molecular structure. These indices include (but are not limited to) the following descriptors: simple and valence path, cluster, path/cluster and chain molecular

connectivity indices [130-132], kappa molecular shape indices [133, 134], topological and electrotopological state indices [135-137], differential connectivity indices, graph's radius and diameter [138], Wiener and Platt indices, Shannon and Bonchev-Trinajstić information indices, counts of different vertices, counts of paths and edges between different kinds of vertices.

Overall, MolConnZ produced over 770 different descriptors. Most of these descriptors characterize chemical structure, but several depend upon the arbitrary numbering of atoms in a molecule and are introduced solely for bookkeeping purposes. In our study, only 644 chemically relevant descriptors were initially calculated and 340 descriptors were eventually used for AmpC beta-lactamase binding dataset after deleting descriptors with zero value or zero variance. MolConnZ descriptors were range-scaled prior to distance calculations since the absolute scales for MolConnZ descriptors can differ by orders of magnitude[139]. Accordingly, our use of range-scaling avoided giving descriptors with significantly higher ranges a disproportional weight upon distance calculations in multidimensional MolConnZ descriptor space.

2.1b.2.5 *k*-Nearest Neighbors (*k*-NN) Classification Method

The *k*NN classification QSAR method [139, 140] is based on the idea that the class that a compound belongs to can be defined by the class membership of its nearest neighbors (i.e., most similar compounds) taking into account weighted similarities between a compound and its nearest neighbors. Since our implementation of *k*NN approach includes variable selection, the similarity is evaluated using only a subset of all descriptors (*nvar*). The similarity is characterized by weighted Euclidean distance between compounds in multidimensional descriptor space. Thus, the class membership of compound *i* can be predicted from the following equation:

$$\hat{y}_i = \sum_{j=1}^{k} \left[\frac{\exp(-d_{ij})}{\sum_{j'=1}^{k} \exp(-d_{ij'})} y_j \right] \qquad (1)$$

where k is the number of nearest neighbors (k = 1 to 5) of compound i, y_j is the class membership of compound j and d_{ij} is the Euclidean distance between compound i and its j^{th} nearest neighbors. In practice, the value of $\hat{y}_i$ is rounded to determine the class membership of compound i:

$$\hat{y}'_i = \text{round}(\hat{y}_i) \qquad (2)$$

The model is internally validated by leave-one-out cross-validation (LOO-CV) where each compound is eliminated from the training set and its class membership is predicted as the class the majority of its k nearest neighbors belongs to. The descriptor set is optimized by simulated annealing approach with the Metropolis-like acceptance criterion to achieve the best CCR value. The CCR is defined as[113]:

$$CCR = 0.5(TP/N_1+TN/N_0) \qquad (3)$$

where N_1 and N_0 are the number of inhibitors and non-binders in the dataset, TP and TN are the number of known inhibitors predicted as inhibitors (true positives) and the number of non-binders predicted as non-binders (true negatives). The statistical significance of the training and test set models is characterized by the LOO-CV CCR_{train} and predictive CCR_{test}, respectively. In summary, the variable selection kNN classification method generates a model with the highest value of CCR that is characterized by the optimal k value, the

number of nearest neighbors, and a subset of selected descriptors. Additional details of this approach can be found elsewhere[139, 141].

2.1b.2.6 Applicability Domain of *k*-NN Models

When developing *k*NN QSAR models, each compound is represented as a point in *M*-dimensional descriptor space (where *M* is the total number of selected descriptors); thus, the molecular similarity between any two molecules can be characterized by the Euclidean distance between their representative points. The Euclidean distance $d_{i,j}$ between two points *i* and *j* (which correspond to compounds *i* and *j)* in *M*-dimensional space can be calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{M} (X_{ik} - X_{jk})^2} \qquad (4)$$

Compounds with the smallest distance between one another are considered to have the highest similarity.

Theoretically, for any compound that can be represented by its MZ descriptors one should be able to predict its class membership using classification *k*NN approach. However, if the distance between the query compound and its *k* nearest neighbors in the training set is large, then the query compound is too dissimilar to the training set compounds, and the prediction of its activity using *k*NN approach appears meaningless. Therefore, a similarity threshold (or model applicability domain) should be introduced to avoid making predictions for compounds, which differ substantially from the training set molecules [19]. The similarity threshold is defined as follows:

$$D_T = \bar{y} + Z\sigma \quad (5)$$

Here, $\bar{y}$ is the average Euclidean distance of the *k* nearest neighbors of each compound within the training set (where the value of k is the same as in predictive *k*NN QSAR models), σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. Typically, we set Z to 0.5, which places the boundary for deciding whether a compound is within or outside of the applicability domain at one-half of the standard deviation. It is important to notice that increasing the value of Z would increase the number of compounds in the external set that are considered within the applicability domain but could decrease the accuracy of prediction due to inclusion of dissimilar nearest neighbors.

2.1b.2.7 Y-randomization Test

Y-randomization test is widely used to ensure model robustness [142]. It includes rebuilding the training set models using randomized activities (Y-vector) of the training set and comparing the resulting model statistics with that for the original test set. It is expected that models built with randomized activities should have significantly lower CCR value for both the training and test sets. In the model building process, it is possible that sometimes, though infrequently, high CCR values may be obtained due to a chance correlation or structural redundancy of the training set. If QSAR models obtained in the Y-randomization test have relatively high LOO-CV CCR_{train} as well as predictive CCR_{test}, it implies that acceptable QSAR models cannot be obtained for the given dataset by the current modeling method. In this study, the Y-randomization test was performed twice for each training/test set splits.

2.1b.2.8 Virtual Screening using *k*-NN Models

As mentioned above, the screening database included 69653 compounds tested by the NCGC against AmpC beta lactamase. The primary HTS screening assay identified 64 "true" hits. Thus, we chose to screen the same database *in silico* using QSAR models as predictors. Only QSAR models that passed both internal and external validation tests were used. For each model we retained its parameters established in the process of external validation, i.e., the number of nearest neighbors *k*, selected descriptors, and Z_{cutoff} value for the applicability domain.

2.1b.3 Results and Discussions

2.1b.3.1 *k*-NN Binary Classification Models

As shown in **Figure 2.1b.2**, the *k*NN QSAR method with variable selection afforded multiple models with optimal accuracy characterized as CCR for both training and test sets. In total, there were 3305 models with both CCR_{train} and CCR_{test} equal or higher than 0.70. Most models with $CCR_{test} \geq 0.70$ also had corresponding $CCR_{train} \geq 0.70$, but the opposite was not always true. The models with high values of both CCR_{train} and CCR_{test} (≥ 0.70) were considered acceptable. 342 predictive models with the highest values of CCR (CCR_{train} and CCR_{test} >= 0.90, red dots in **Figure 2.1b.2**) were selected for consensus prediction. **Table 2.1b.1** summarizes the detailed confusion matrix and statistical parameters for the best *k*NN binary classification models. The CCR_{train} and CCR_{test} were found to be as high as 0.91 and 1.00, respectively, which implies that the models could identify correctly all 18 nonbinders and 9 out of 11 inhibitors (SE = 0.82, SP = 1.00, EN(1) = 2.00, and EN(0) = 1.69) in the training set and all binders and non-binders in the test set. This remarkably high internal accuracy and the large number of acceptable models imply that the *k*NN classification

method was generally successful in correctly distinguishing binders *vs*. decoys using MolConnZ chemical descriptors of compounds only.

2.1b.3.2 QSAR Model Validations

In addition to the internal validation of *k*NN models using test sets, Y-randomization and external validation are the critical steps of the entire QSAR workflow (**Figure 2.1b.1**). Only models that have been validated by these two steps can be utilized for external prediction and database mining [19].

Y-randomization Test

In Y-randomization test, the binary annotations of AmpC beta-lactamase as inhibitors or non-binders were randomly shuffled and *k*NN classification models were built with the same parameter setting. The test was performed twice and both runs of Y-randomization tests showed that there were relatively small numbers of 330 and 429 models having both CCR_{train} and CCR_{test} higher than 0.70. However, there were no models with both CCR value higher than 0.90. It implied that the *k*NN models obtained with real binding affinities and CCR greater than 0.90 are robust.

External Validation

Two datasets were employed for external validation, i.e. the 10 compounds randomly excluded from modeling sets and 50 non-binders, which were relatively dissimilar in their structure from the 21 inhibitors in the original dataset. Consensus predictions were carried out using 342 predictive models with CCR_{train} and CCR_{test} greater than 0.9 under different Z value cutoffs (Z = 0.5 ~ 3.0, **Table 2.1b.2**). The prediction accuracy for the 10-compound external validation set was 100% for both 5 inhibitors and 5 non-binders under Z_{cutoff} = 0.5,

leading to CCR = 1.00, SE = 1.00, SP = 1.00, EN(1) = 2.00, and EN(0) = 2.00. The accuracy of prediction for 50 non-binders was also high, ranging from CCR = 0.87 under Z_{cutoff} = 0.5 to CCR = 0.86 under Z_{cutoff} = 3.0 (**Table 2.1b.2**). Because of the applicability domain inherent to individual *k*NN QSAR models, the consensus prediction usually can not cover the whole dataset. By increasing the Z_{cutoff} from 0.5 to 3.0, the prediction coverage for 50 non-binders increased from 94% to 98% whereas the prediction accuracy decreased. **Figure 2.1b.3** shows the consensus scores and the coverage of predictive models for each of the 50 non-binders. The consensus score, in terms of the average class number in classification QSAR, was calculated by the fraction of models that predicted a compound as non-binder over the total number of models used for prediction plus 1. Under Z_{cutoff} = 0.5, six falsely predicted inhibitors (average class number < 1.5) were within the applicability domain of only 70 models (i.e., approximately 20% of all models), i.e., the model coverage was as low as 20%. In general, the prediction with such a low coverage is viewed as of low confidence level. The higher Z_{cutoff} significantly raised the model coverage for both inhibitor and non-binder prediction because of the extended applicability domain for individual models. In **Figures 2.1b.3B** and **2.1b.3C**, the model coverage for predicting inhibitors jumped up to 53% for Z_{cutoff} = 1.5 and up to 94% for Z_{cutoff} = 3.0. However, the prediction with extended applicability domain for consensus models also comes with lower confidence level. Generally speaking, in order to have the reliable and accurate prediction, one has to have the broader model coverage and a smaller Z_{cutoff} value.

In summary, 342 models with both CCR_{train} and CCR_{test} equal to or greater than 0.90 could be applied for consensus prediction and database mining. The models chosen for the

prediction had relatively small Z_{cutoff} (= 0.5) and relatively broad coverage for compounds in external datasets (>= 50%).

2.1b.3.3 External Prediction

We used models built from 41 AmpC inhibitor/nonbinder dataset to verify the 64 "actives" from AID 584 and AID 585 screening. Under Z_{cutoff} = 0.5, we could only generate predictions for 25 compounds out of 64 "actives" whereas the remaining compounds were found to be outside of the applicability domain. As shown in **Table 2.1b.2**, five out of these 25 compounds were predicted as true inhibitors. However, the predictions were based on only two models (out of 342 models with both CCR_{train} and CCR_{test} higher than 0.90, cf. **Figure 2.1b.4A**). Thus, the coverage for both compounds and consensus models was extremely low and as a result these predictions should not be viewed as reliable. Even under higher Z_{cutoff} = 3.0, the model coverage was still low such that "actives" were predicted by only 110 models (32% of all models, cf. **Figure 2.1b.4C**). Furthermore, the formal prediction accuracy (assuming that the 64 hits were true inhibitors) was extremely low, e.g. CCR = 0.20 (Z_{cutoff} = 0.5), CCR = 0.10 (Z_{cutoff} = 1.5) and CCR = 0.15 (Z_{cutoff} = 3.0) (**Table 2.1b.2**). Thus, based on our modeling results none of the 64 compounds in the NCGC set was predicted reliably as a non-covalent and reversible inhibitor.

Notably, the independent experimental verification of those 64 "actives" hits appears to confirm the results of our consensus prediction based on recent results obtained in Dr. B. Shoichet's lab. These studies [122] have shown that 25 of these active compounds are beta-lactam-based irreversible inhibitors of beta-lactamase. Five to ten additional actives are believed to be aggregators. The data on the remaining 35 compounds have not been

confirmed yet but preliminary data indicate that none of them act as true reversible inhibitors of beta-lactamase (Dr. Shoichet, personal communications). These recent results confirm that our models are both accurate and robust.

2.1b.3.4 Descriptor Interpretation

A summary of descriptors ranked as top 20 based on their frequency of occurrences in 342 consensus models are given in **Table 2.1b.3**. The frequency of occurrence is defined as percentage of models where a descriptor is present. For instance, the highest frequency of 32.2% means that a particular descriptor type occurs in about 110 out of the total of 342 models. The descriptor class and the structural illustration of individual descriptor types are shown in this Table as well. It should be noted that molecular connectivity descriptors are predominant in all models, i.e., over 50% of descriptors in 20 top-ranked most frequent descriptor types belong to this class. The remaining descriptor types are mostly related to class of electrotopological state (E-state) indices which reflect the electronic environment of each atom due to its intrinsic electronic properties and the influence of other atoms in the molecule. By mapping the frequent descriptors to the inhibitors and non-binders in the dataset, the sulfonamide group was found to be a common feature in both inhibitors and non-binders. Importantly, all the nitrile groups could only be found in the structure of non-binders. Thus, conventional structure based scoring functions appear to be insensitive to (the presence or absence of) this group in chemical structures. This result illustrates a potential power of QSAR models in informing conventional scoring functions of their possible deficiencies that probably could be corrected with ease.

2.1b.3.5 Virtual Screening Using Predictive QSAR Models

Instead of using only one single and best model for virtual screening, the consensus prediction approach was applied that relies on averaging predictions from all qualified models, i.e. 342 models with both CCR_{train} and CCR_{test} equal to or greater than 0.90. The complete modeling set (i.e., including training and test sets) was used for the prediction using each model as opposed to using only the corresponding training set. Initially, as many as 4565 compounds in the NCGC dataset of 69653 compounds were predicted as inhibitors by at least one of 342 models. To narrow the hit list and obtain the higher confidence level for each prediction, we took both the consensus score (average class number) and model coverage into account. In particular, only the hits with average class number between 1.0 and 1.2 and the model coverage over 50% (171 out of 342 models) were selected (**Figure 2.1b.5**). Furthermore, we restricted ourselves to the most conservative applicability domain for each model using Z_{cutoff} = 0.5. We found that there were only 15 compounds that satisfied both criteria (**Table 2.1b.4**).

We have clustered these 15 compounds together with 16 competitive inhibitors from the training set using tools available in PubChem [143]. Each compound was represented by a fingerprint of 881 substructure keys, indicating the presence or absence of a particular chemical substructure. The pairwise similarity between compounds was measured by the Tanimoto coefficients (TC), which were used for hierarchical clustering of hits. The most chemically different pair of structures had TC = 0.522 (**Figure 2.1b.6**). Several structural classes were observed depending on the TC thresholds, e.g. there were four clusters at TC = 0.70. Notably, many of the 15 computational hits were found to be structurally similar to inhibitors used in model building. There were five hits that were highly similar (TC >= 0.90;

CID: 39854, 665205, 699751, 793725 and 699907) to the competitive inhibitor 11771345. More often than not, hit compound 647810 was in close proximity to inhibitor 11347033 with the Tanimoto coefficients over 0.90. Several computational hits were selected for the experimental validation in Dr. Shoichet's laboratory as potential AmpC beta-lactamase inhibitors.

We should emphasize that our model validation is a critical inherent feature of our QSAR modeling workflow. This issue of model validation has been given a lot of attention by the QSAR research community [144]. Until recently, most practitioners merely presumed that internally cross-validated models built from available training set data should be externally predictive. We and others have demonstrated that internal validation techniques such as leave-one-out (LOO) or even leave-many-out (LMO) cross-validation applied to the training set is insufficient to ensure the external predictive power of QSAR models [18, 19]. Thus, we used two external validation sets in this study as well as the Y-randomization test to ensure the robustness and predictive power of *k*NN models. Needless to say, the use of externally validated models and applicability domains is especially critical when the models are employed in virtual screening.

Another important feature of many current biomolecular datasets, especially generated as a results of HTS campaigns is the imbalance between "actives" and "inactives", obviously in favor of inactives. For example, the hit rates in assays deposited in PubChem by the NIH screening centers forming the Molecular Library Screening Center Network (MLSCN) are very low, in most cases not exceeding 0.5% [145]. The imbalanced datasets pose a significant problem for classification QSAR modeling because models that predict correctly the same fraction of objects in each class will have different objective function

values. To circumvent this problem in this study, we conducted the similarity search between the members of the underrepresented class (inhibitors) vs. another one (non-binders). A subset of the original dataset that was relatively balanced (2:3) was formed and utilized for model building. The 50 non-binders that were less similar to binders were retained as one of the external validation datasets. The classification models built for the balanced subset were shown to predict compounds in this external dataset as non-binders with very high accuracy. Among the 47 non-binders (3 were outside of the applicability domain), 41 were accurately annotated by consensus prediction (CCR = 0.87, cf. **Table 2.1b.2**). The success of this strategy suggests that it could be applied to the analysis of many imbalanced datasets.

2.1b.3.6 Experimental Validation

Of the 15 computational hits from mining the NCGC AmpC screening library, five compounds were selected based on their chemical similarity (measured by Euclidean distance in the MZ descriptor space) to the 21 inhibitors and commercial availability. We should stress that binary QSAR models were used for prediction so no quantitative estimate of binding affinity could be made. All five hits (CID: 647810, 665205, 699751, 699907 and 2980565; **Table 2.1b.4**) did show the inhibitory activities at millimolar level at the single concentration. Among them, compound 699751 had the highest inhibitory activity at 0.7 mM. For this compound, a full dose-response curve was obtained and the inhibition constant, K_i, was calculated by the Cheng-Prusoff equation using the IC_{50} and K_d, the dissociation constant of AmpC for the substrate measured in a separate assay. Thus, compound 699751 yielded the K_i and K_d value of 135 and 18 □M, respectively (**Figure 2.1b.7**). In summary, the above results did prove the predictive power of our binary *k*NN classification QSAR models built for AmpC beta-lactamase inhibitors. These studies illustrate that the validated

QSAR workflow, as employed in this paper, could be used as a general tool for identifying promising hits by the means of virtual screening of chemical libraries.

2.1b.4 Conclusions

Our studies demonstrate that binary *k*NN classification QSAR models built with MolconnZ descriptors can accurately differentiate true AmpC beta-lactamase inhibitors from non-binding decoys. A special QSAR modeling scheme was employed for this imbalanced dataset and the models were rigorously validated using both internal (multiple training/test set divisions and Y-randomization) as well as external (two external validation sets) validation approaches. We have demonstrated that this strategy afforded multiple QSAR models with high internal and external predictive power. As part of our QSAR modeling workflow, the predictors were further utilized for mining the NCGC dataset (69653 compounds tested for AmpC beta-lactamase binding). We found that our validated models disagreed with the experimental annotation of 64 compounds as AmpC binders as reported in PubChem BioAssays AID584 [123] and AID585 [124]. Interestingly, our negative predictions for these compounds appear to be in agreement with the preliminary results of the confirmatory secondary assays conducted in B. Shoichet's lab (B. Shoichet, personal communications). On the other hand, our models used in the most conservative way (i.e., in consensus fashion and with the strictest applicability domain criteria) did identify 15 putative AmpC inhibitors among compounds annotated as experimental non-binders in the NCGC assays reported in PubChem. Five of them showed inhibition activities at the millimolar concentration, and one compound (compound 699751) was found to have the highest Ki of 135□M. The results of our studies suggest that at least in some cases when a sufficient amount of data on true binders vs. nonbinding compounds is available simple QSAR modeling approaches could be

used successfully to complement (and possibly educate based on QSAR model interpretation) the conventional scoring functions used in three-dimensional docking studies. Furthermore, as we have demonstrated in this paper, QSAR models can be successfully used not only to discriminate binders vs. binding decoys but most importantly, for finding promising hits by the means of virtual screening of chemical libraries.

Figures for Chapter 2.1b

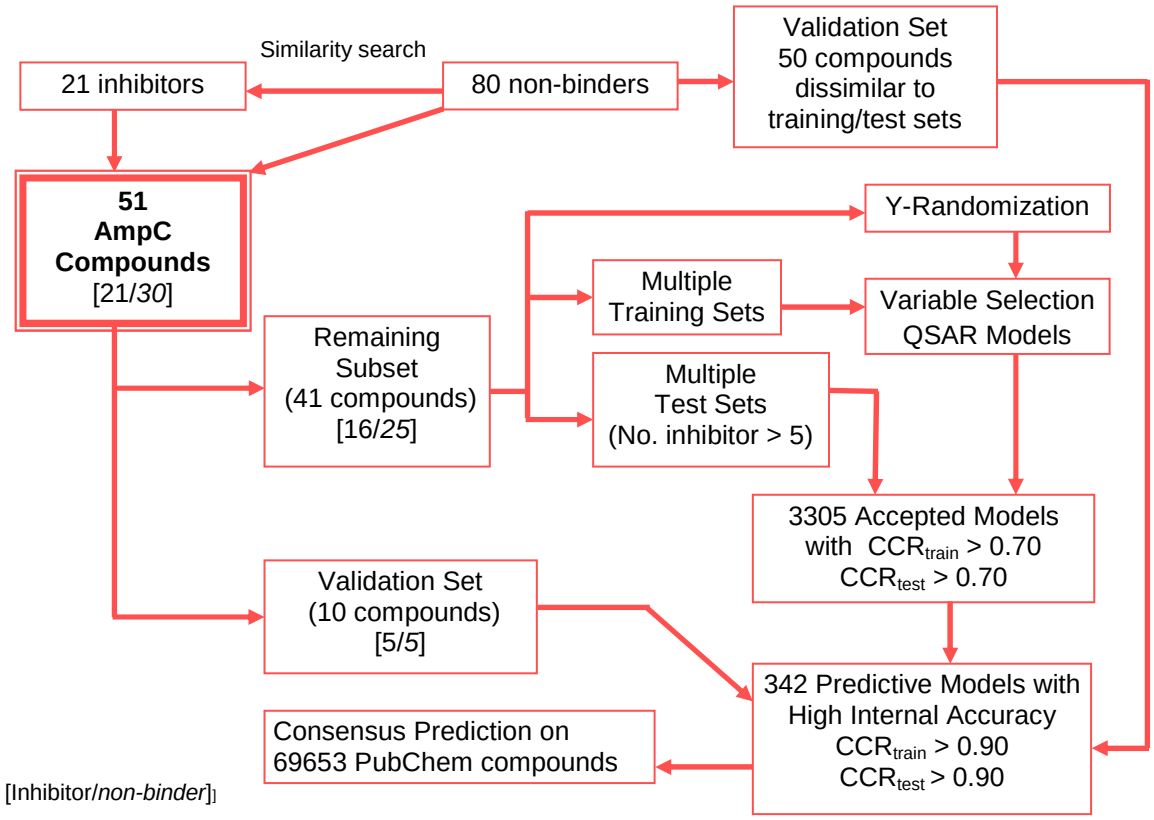


Figure 2.1b.1: The workflow of QSAR model building, validation, and virtual screening as applied to the AmpC beta-lactamase dataset of 21 inhibitors and 80 non-binding decoys.

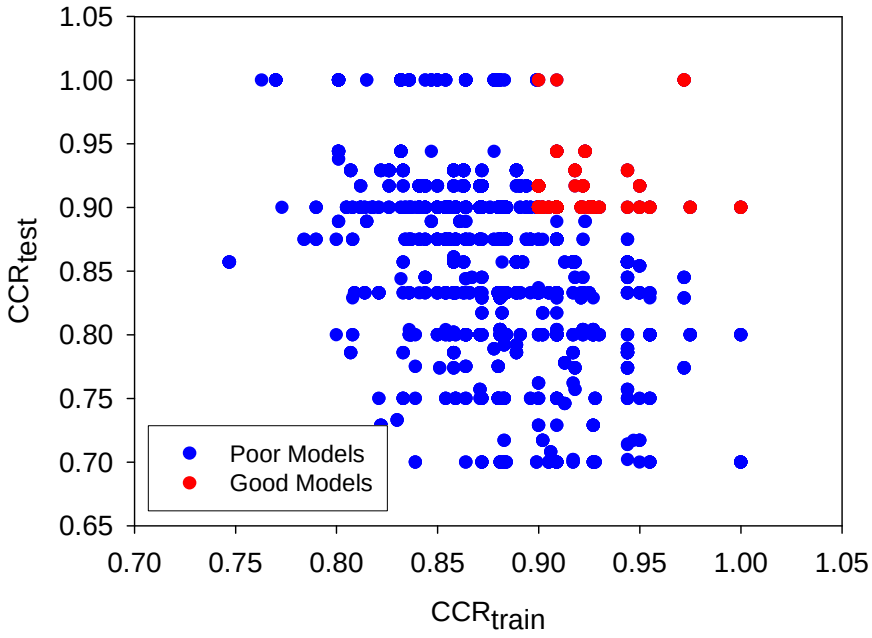


Figure 2.1b.2: The plot of *k*NN classification QSAR model accuracy for test (CCR_{test}) vs. training (CCR_{train}) sets for AmpC beta-lactamase dataset.

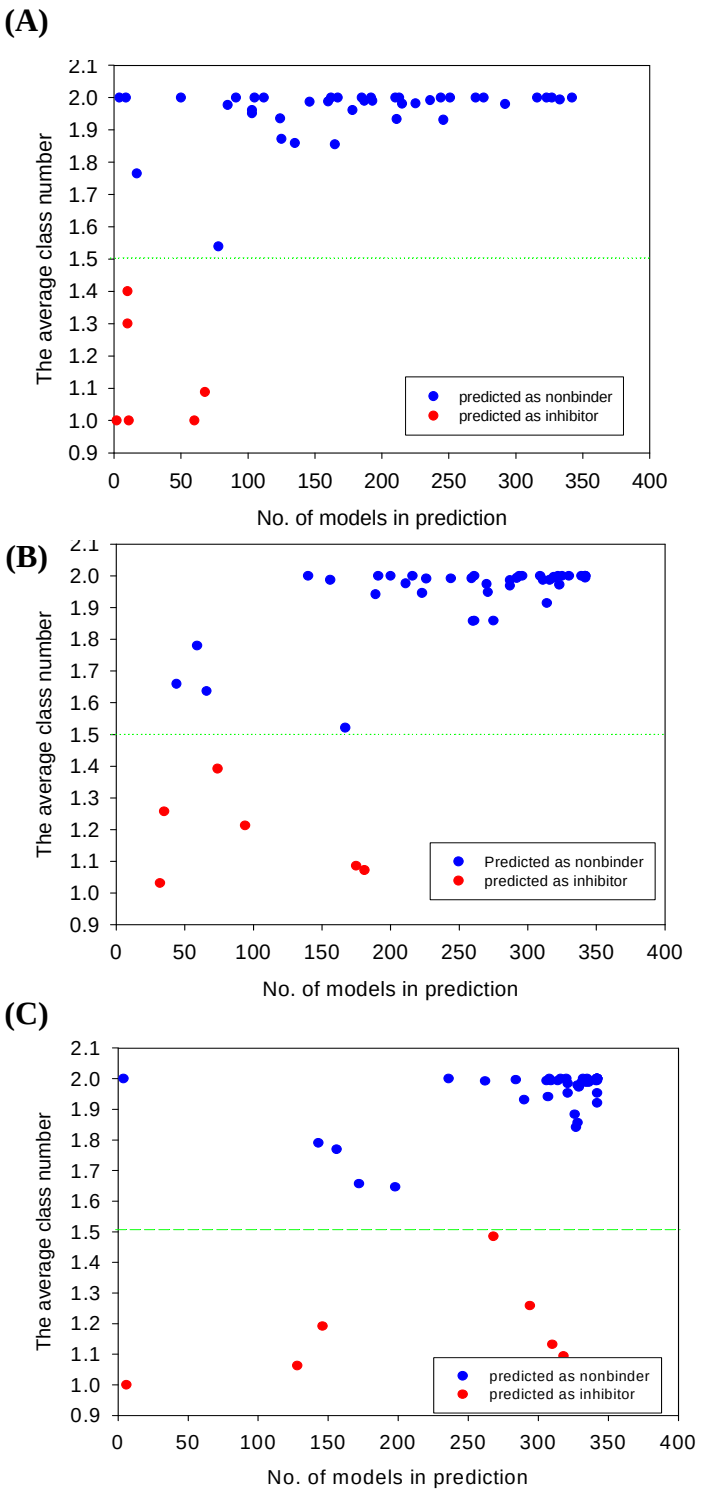
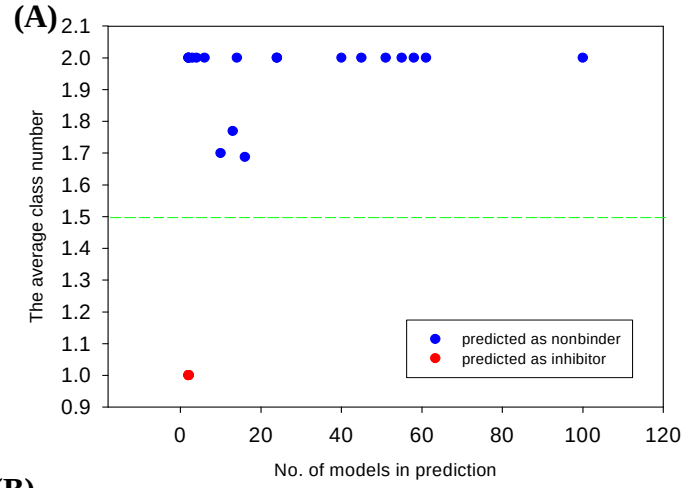


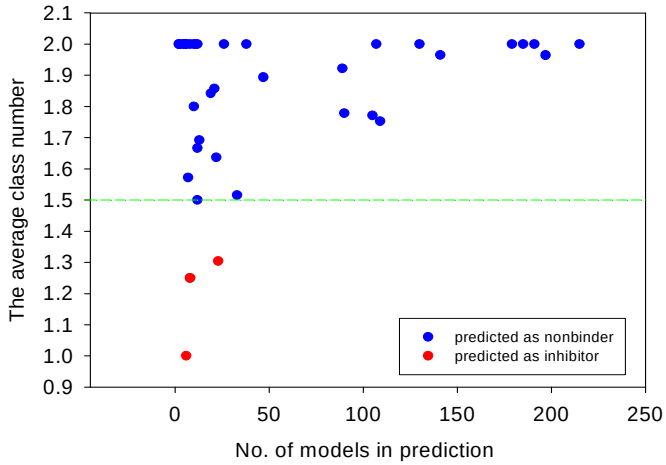
Figure 2.1b.3: The consensus scores and the coverage of predictive models for the 50 non-binding decoys dissimilar to the modeling dataset.

Three Z cutoff values were used: A. Z_{cutoff} = 0.5; B. Z_{cutoff} = 1.5; C. Z_{cutoff} = 3.0.

(A)



(B)



(C)

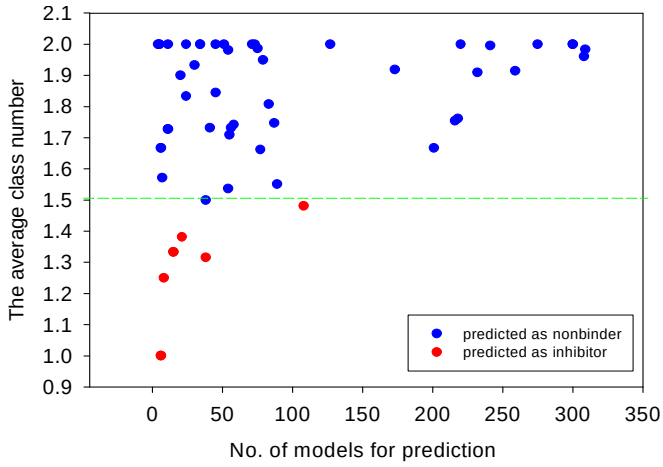


Figure 2.1b.4: The consensus scores and the coverage of predictive models for the 64 HTS hits identified from the primary HTS screening assays reported in PubChem.

Three Z cutoff values were used: A. Z_{cutoff} = 0.5; B. Z_{cutoff} = 1.5; C. Z_{cutoff} = 3.0.

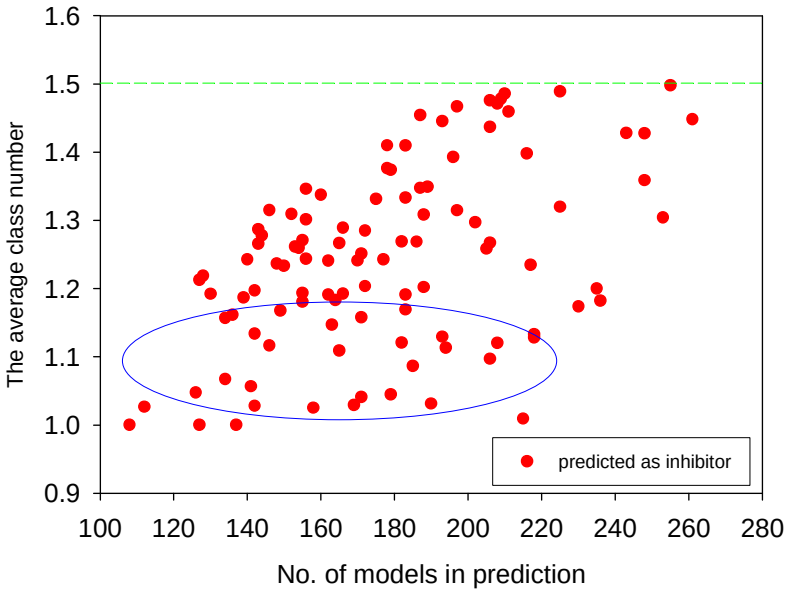


Figure 2.1b.5: The consensus scores and the coverage of predictive models for the mining hits in the NCGC database (Z_{cutoff} = 0.5).

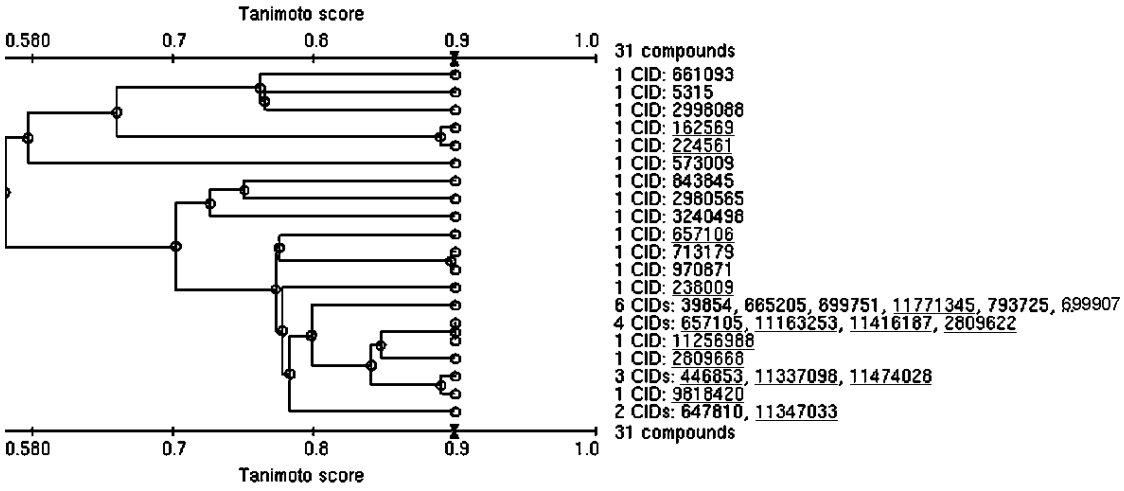
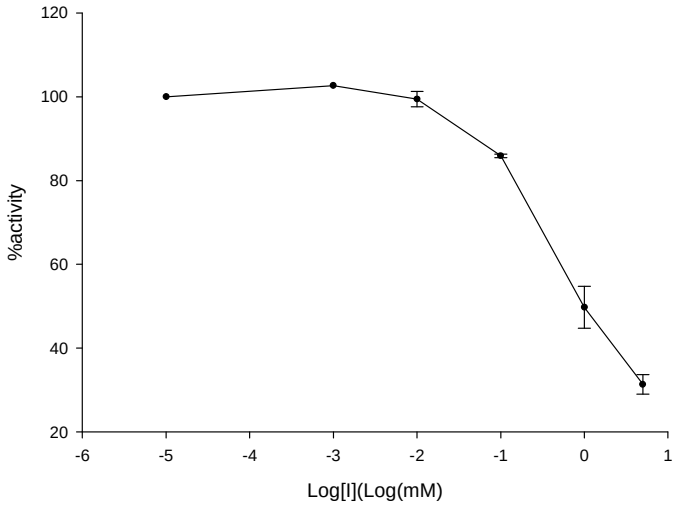
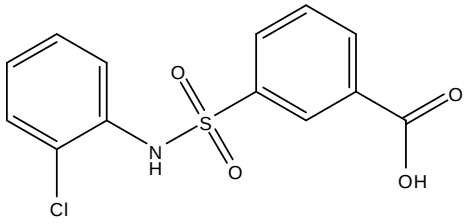


Figure 2.1b.6: The structural clustering of 15 mining hits from NCGC database combined with 16 AmpC beta-lactamase competitive inhibitors (underlined) based on the Tanimoto score.

The computations were carried out at the PubChem server.



CID: 699751



Kd = 18µM, Ki = 135µM

Figure 2.1b.7: The full dose response curve for compound 699751.

The experimental studies were conducted in B. Shoichet's laboratory.

Tables for Chapter 2.1b

Table 2.1b.1: Ten best *k*NN QSAR classification models with highest CCR values for all test sets using Molconnz descriptors.

Model No.	Nearest Neighbors No.	Confusion Matrix						Statistics for the Models				
		CCRtrain	N(1)	N(2)	TP	TN	FP	FN	SE	SP	EN(1)	EN(2)	CCRtest
1	5	0.91	11	18	9	18	0	2	0.82	1.00	2.00	1.69	1.00
2	5	0.90	10	18	8	18	0	2	0.80	1.00	2.00	1.67	1.00
3	1	0.97	11	18	11	17	1	0	1.00	0.94	1.89	2.00	1.00
4	5	0.91	11	16	9	16	0	2	0.82	1.00	2.00	1.69	0.94
5	4	0.92	11	16	10	15	1	1	0.91	0.94	1.87	1.82	0.94
6	4	0.92	9	19	8	18	1	1	0.89	0.95	1.89	1.79	0.93
7	1	0.94	10	18	10	16	2	0	1.00	0.89	1.80	2.00	0.93
8	5	0.92	10	19	9	17	1	1	0.90	0.89	1.89	1.80	0.92
9	5	0.92	9	19	8	18	1	1	0.89	0.95	1.89	1.79	0.92
10	5	0.90	10	17	8	17	0	2	0.80	1.00	2.00	1.67	0.92

N(1) = number of inhibitors, N(2) = number of non-binders, TP = true positive (inhibitors predicted as inhibitors), FP = false positives (non-binders predicted as inhibitors), FN = false negatives (inhibitors predicted as non-binders), TN = true negative (non-binders predicted as non-binders), SE = sensitivity = TP/N(1), SP = specificity = TN/N(2), EN - the normalized enrichment, EN(1) = (2TP * N(2))/(TP * N(2) + FP * N(1)), EN(2) = (2TN * N(1))/(TN * N(1) + FN * N(2)), and CCR = correct classification rate.

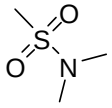
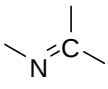
Table 2.1b.2: Consensus predictions under different Z value cutoffs for two external validation sets, the randomly-excluded 10 compounds from modeling sets and 50 non-binders which were dissimilar in structure to 21 inhibitors in the original dataset.

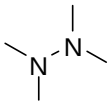
External Validation Sets	Z_{cutoff}	Prediction CCR	Confusion Matrix						Statistics			
			N(1)[a]	N(2)[a]	TP	TN	FP	FN	SE	SP	EN(1)	EN(2)
10 randomly-excluded compounds	0.5	1.00	5	5	5	5	0	0	1.00	1.00	2.00	2.00
50 non-binders	0.5	0.87	0	47	0	41	6	0	N/A	0.87	N/A	N/A
	1.5	0.87	0	47	0	41	6	0	N/A	0.87	N/A	N/A
	3.0	0.86	0	49	0	42	7	0	N/A	0.86	N/A	N/A
64 HTS 'hits'	0.5	0.20	25	0	5	0	0	20	0.20	N/A	N/A	N/A
	1.5	0.10	41	0	4	0	0	37	0.10	N/A	N/A	N/A
	3.0	0.15	55	0	8	0	0	47	0.15	N/A	N/A	N/A

N(1) = number of inhibitors, N(2) = number of non-binders, TP = true positive (inhibitors predicted as inhibitors), FP = false positives (non-binders predicted as inhibitors), FN = false negatives (inhibitors predicted as non-binders), TN = true negative (non-binders predicted as non-binders), SE = sensitivity = TP/N(1), SP = specificity = TN/N(2), EN - the normalized enrichment, EN(1) = (2TP * N(2))/(TP * N(2) + FP * N(1)), EN(2) = (2TN * N(1))/(TN * N(1) + FN * N(2)), and CCR = correct classification rate.

[a]Many N(1) inhibitors of 64 HTS 'hits' and N(2) non-binders of 50 non-binders were out of application domain of all consensus models, thus having no prediction. Only data for compounds found within the AD were used for statistical summaries.

Table 2.1b.3: The 20 most frequent MolConnZ descriptors found in acceptable *k*NN QSAR models.

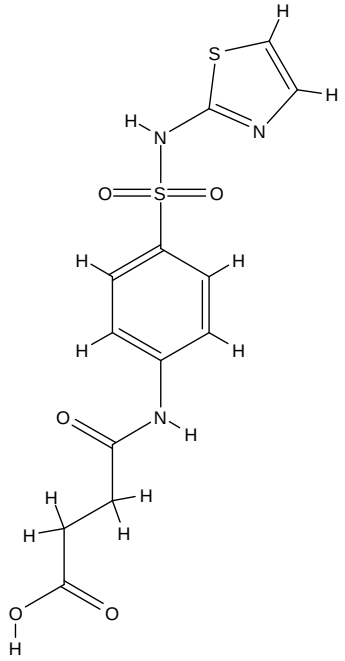
Rank[a]	Descriptor ID	Frequency[b]	Descriptor Class	Illustration
1	**nHCsatu**	32.2	atom-type counts	CH_n (unsaturated)
2	**Hsulfonamide**	28.4	group-type Hydrogen E-State values	
3	**nnitrile**	27.5	group-type counts	—C≡N
4	**Hmin**	27.2	minimum H E-State	
5	**naaO**	26.3	atom-type counts	:O: (aromatic)
6	**naaS**	26.3	atom-type counts	aSa (aromatic)
7	**SaaCH**	26.0	atom-type EState sums	:CH:
8	**n3Pad24**	26.0	vertex alpha-delta counts	
9	**SssCH2**	26.0	atom-type EState sums	$-CH_2-$
10	**SHBint5**	25.4	internal H-Bond counts and EStates	
11	**Xvch5**	24.3	valence cluster/chain Chi indices	
12	**n2Pag23**	24.3	vertex alpha-gamma counts	
13	**IDW**	24.0	Bonchev-Trinajstic information indices	
14	**htets2**	23.7	total topological state indices based on H E-State indices	
15	**nimine**	23.7	group-type counts	
16	**ndsCH**	23.4	atom-type counts	=CH-
17	**IDC**	23.4	Bonchev-Trinajstic information indices	
18	**tets3**	23.1	total topological state indices based on E-State indices	

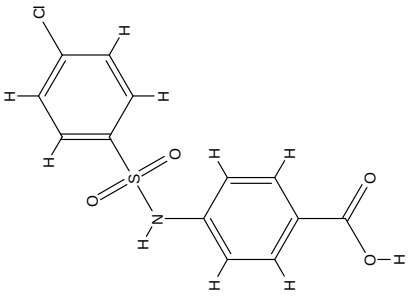
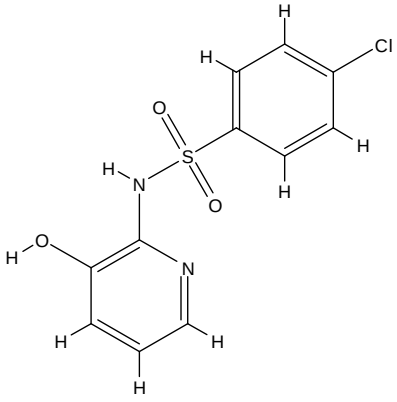
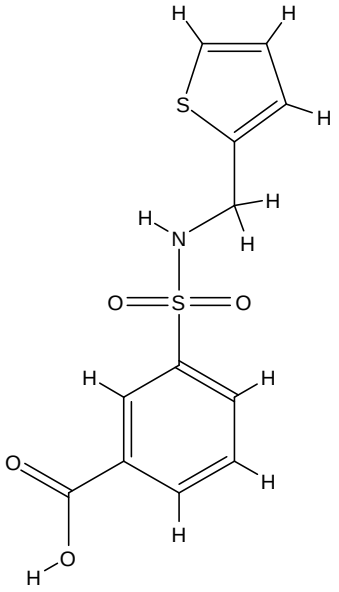
19	**n3Pad13**	23.1	vertex alpha-delta counts	
20	**nhydrazine**	22.8	group-type counts	

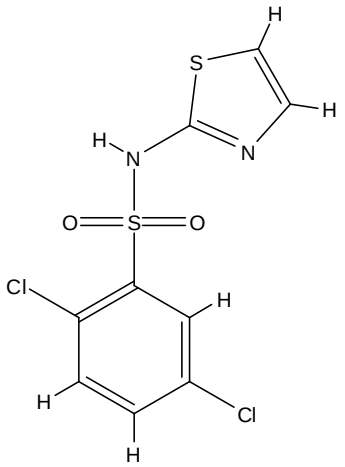
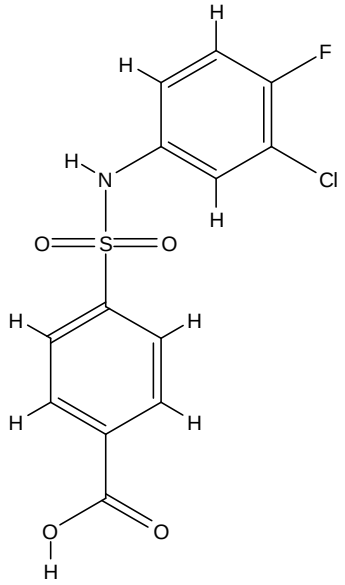
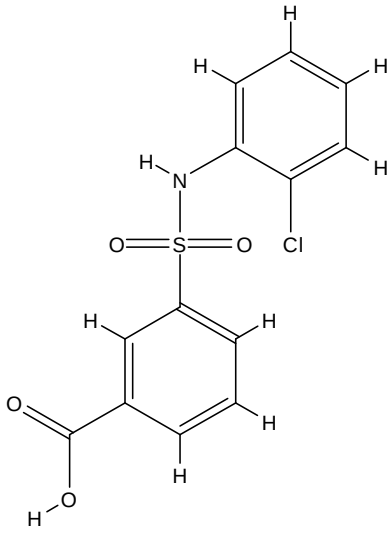
[a] *k*NN rank is based on the frequency of each descriptor occurred.

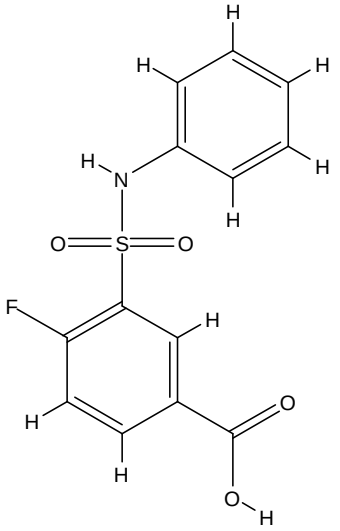
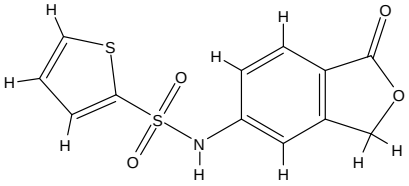
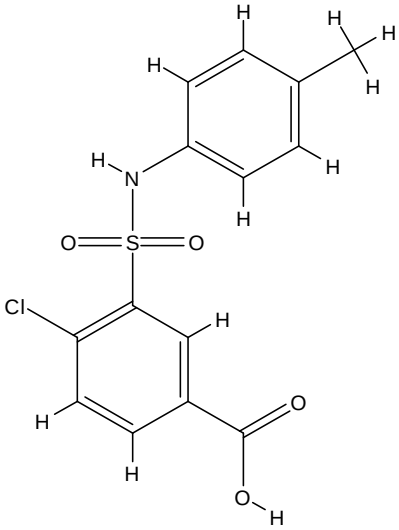
[b] Frequency is the number of times each descriptor occurred in 342 validated models.

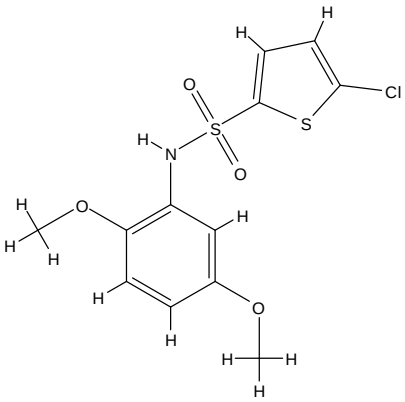
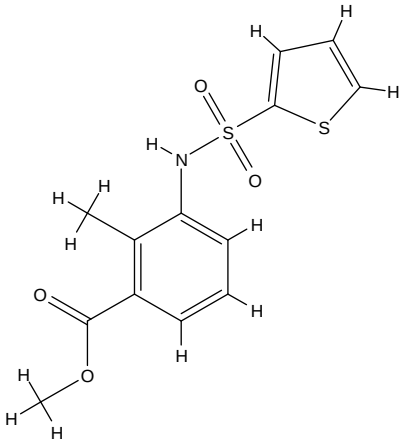
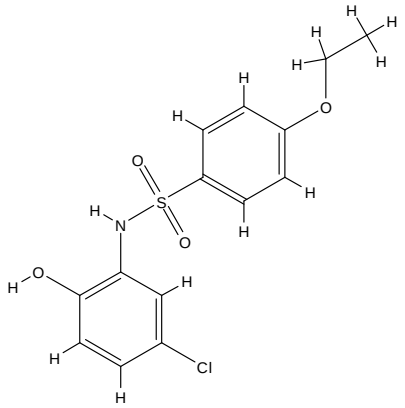
Table 2.1b.4: The fifteen computational hits predicted as AmpC beta-lactamase inhibitors as a result of mining the NCGC AmpC screening library.

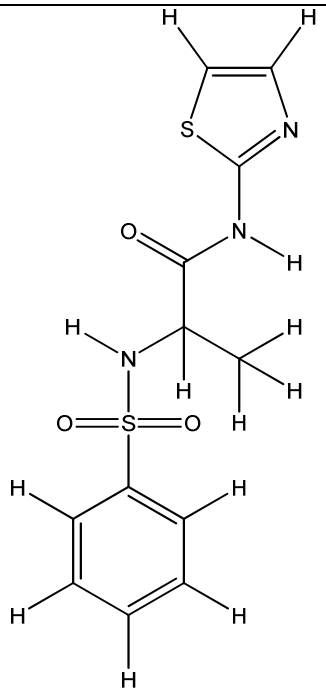
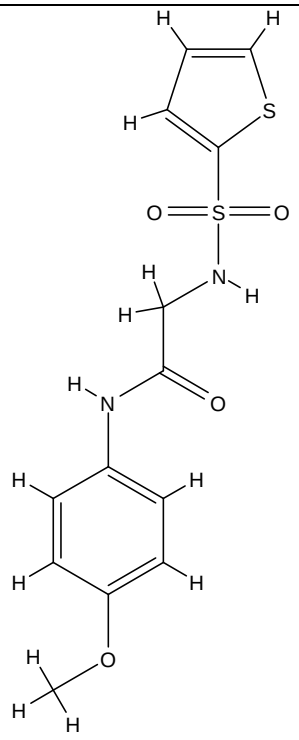
Structure	Serial No.	PubChem CID	No. of models predicted as inhibitor	No. of models predicted as non-binder	No. of models in prediction	Average class num.	Exp. IC_{50} (mM)[a]
	1	5315	213	2	215	1.01	Unknown

	2	39854	186	20	206	1.10	Unknown
	3	573009	190	40	230	1.17	Unknown
	4	647810	193	43	236	1.18	3.0

	5	661093	172	22	194	1.11	Unknown
	6	665205	171	8	179	1.04	9.0
	7	699751	189	29	218	1.13	0.7[b]

	8	699907	184	6	190	1.03	1.8
	9	713179	169	16	185	1.09	Unknown
	10	793725	168	25	193	1.13	Unknown

	11	843845	152	31	183	1.17	Unknown
	12	970871	183	25	208	1.12	Unknown
	13	2980565	190	28	218	1.13	7.0

	14	2998088	160	22	182	1.12	Unknown
	15	3240498	148	35	183	1.19	Unknown

Chapter 3 Development of Quantitative Structure-Binding Affinity Relationship Models (QSBAR) Using Protein-Ligand Interface Descriptors Based on Conceptual Density Function Theory (DFT) and the Application to Community Structural-Activity Resources (CSAR) Data Sets

3.1 Introduction

Predicting binding affinity of protein-ligand complexes, either relative or absolute, plays an essential role in structure-based drug design/discovery. In all structure-based drug design methods, if the experimental protein-ligand structural information is unknown, docking and *scoring functions* are applied jointly to generate putative poses for binding affinity prediction. Since 1980s, many scoring functions have been developed and been critically assessed recently.[60, 146, 147] These studies demonstrate that correctly predicting binding affinity of compounds by traditional docking/scoring functions is still fairly challenging. The squared correlation coefficient (R^2) between experimental and predicted binding affinities of several popular scoring functions is in the range of 0.3 to 0.4 when predicting binding affinity of static x-ray structures,[60] not even to mention the R^2 value when conducting cross-docking calculations where protein induced-fit effects are encountered. The result suggests that, for other than some computationally intensive approaches (e.g., free energy perturbation[148] or linear interaction energy[58]) recently being developed for target-specific lead optimization in structure-based drug design,[58, 59] the improvement of current scoring functions for generic high-throughput molecular docking is also needed.

In general, scoring functions can be divided into three major classes:[27] force-field-based, empirical, and knowledge-based methods. The force-field-based scoring function relies on explicitly computed electrostatic and van der Waals interaction energies (i.e., enthalpic effects) between the ligand and the protein based on a molecular force field. The empirical scoring function is defined as the sum of individual uncorrelated energy terms whose coefficients are optimized from regression analysis by fitting the experimental data such as binding energies/affinities. The knowledge-based scoring function is designed based on various statistical parameters derived from x-ray crystal structures that could reflect the interactions between a ligand and its receptor depending on their molecular environment. Compared with force-field-based and empirical scoring functions, knowledge-based scoring functions could implicitly capture the binding effects that are difficult to model and specify (e.g., entropy and solvation) by analyzing the statistics of atomic contacts based on a large number of experimentally determined protein-ligand structures. However, the score from knowledge-based scoring functions usually corresponds to the sums of pair interactions or other binding effects, which are indirectly associated with the absolute binding affinity.

The tuning of above scoring functions relies on the availability of structural information of protein-ligand complexes. In contrast, typical cheminformatics approaches (e.g., QSAR modeling) usually do not require the protein-ligand structural information and absolute binding affinity of ligands can be predicted as a function of their chemical descriptors. Recently, a hybrid scoring function incorporating cheminformatics concepts into conventional scoring functions was developed in our lab.[16] This scoring function consists of quantitative structure-binding affinity relationship (QSBAR) models derived

from 264 x-ray protein-ligand complexes (the old ENTess data set) with known binding affinity using novel protein-ligand interfacial geometrical properties, called ENTess descriptors. The ENTess descriptors are generated based on the tetrahedra resulting from Delaunay tessellation (Tess), characterizing the protein-ligand interface by means of Pauling electronegativity (EN) values. The output of ENTess scoring function can be directly related to absolute binding affinities and can implicitly take into account binding effects that are difficult to specify, combining the merit of both empirical and knowledge-based scoring function. However, the performance of ENTess scoring function in practical virtual screening is limited (data not shown). One of the possible reasons could be the limitation in applicability domain of ENTess models. In the following study, we report the study managing to improve the ENTess scoring function.

Taking the advantage of rapidly increasing number of x-ray protein-ligand complexes with experimentally determined binding affinity, firstly we extend the previous study by including large number of structurally diverse protein-ligand complexes into model building and validation. Those high-quality complexes for modeling are acquired from PDBbind database[149, 150] and Community Structural-Activity Resources (CSAR),[151] both of which are public available. Secondly, we also incorporate theoretically more rigorous values (i.e., conceptual DFT atomic properties[152]) as well as protein-ligand pairwise distances within interfacial tetrahedra into descriptor generation. Finally, in addition to the applicability domain of respective eligible models, i.e., *model* applicability domain, we only predict the protein-ligand complexes which are similar to the complexes of the modeling set in the entire descriptor space (i.e., those within *global* applicability domain).

With above strategies, we predict the binding affinity of x-ray protein-ligand complexes from two data sets (Set1 and Set2) provided in the 2010 CSAR exercise. CSAR 2010 exercise has involved researchers developing their various scoring methodologies in the hope of improving current scoring methods. We achieve comparable prediction accuracy (R^2: 0.57) to the best in the CSAR exercise (R^2: 0.58 by July 2010) using the updated QSBAR models.

3.2 Methods

3.2.1 Data Sets

PDBbind version 2007

We employ the x-ray protein-ligand complexes as well as their ligands' binding affinity (BA) data collected from PDBbind (version 2007) database in our QSBAR study. PDBbind database[149, 150] is a collection of x-ray protein-ligand complexes with experimentally measured binding affinity data (IC_{50}, K_i, or K_d). In total, there are 3124 complexes (a.k.a. the "general" set) included in the version 2007 of PDBbind database. From the complexes of the general set, 1300 complexes are culled to form the "refined" set considered as high-quality (e.g., resolution $\geq$ 2.5Å, non-covalent protein-ligand binding, and experimentally determined K_i or K_d values). The BLAST sequence clustering using 90% similarity threshold is conducted on 1300 protein sequences of refined set, resulting in 70 clusters (families). Three representative complexes with highest, medium, and lowest BA from each of the 70 clusters are collected to form the "core" set. The core set is designed to provide a diverse and non-redundant sampling of the refined set. The relationships between general set, refined set, and core set are presented in **Figure 3.1**.

The proteins in the core set are employed as probes to search for the structurally similar proteins in the refined set in order to construct a new QSBAR modeling set with increased size and diversity (compared with the old data set of 264 protein-ligand complexes, see Introduction). For each protein binding site in the refined set, the protein descriptors[153] are calculated and the similarity threshold is defined using Euclidean distance in the protein descriptor space (i.e., threshold = $\langle d \rangle + 0.5\sigma$, where σ is the standard deviation and $\langle d \rangle$ is the average of distances between each data point in the core set and its nearest data point in the refined set). In total, 455 complexes with protein binding sites similar to those of query proteins are selected, along with 210 complexes in the core set, forming our new data set for QSBAR modeling (665 complexes). The BA range of these 665 complexes is from 1.36 (1qpb.pdb) to 13.96 (7cpa.pdb) and the number of protein families based on 90% sequence similarity clustering is 101 clusters. The new data set is almost three fold larger than the old data set and has more families (101 *versus* 83). The comparison between old and new data set is shown in **Table 3.1**.

In the 2010 CSAR exercise, the researchers are asked to predict the binding affinity of protein-ligand complexes in Set1 and Set2 using their own scoring methods which are trained without CSAR data sets or can be tuned based on either of Set1 or Set2 data set.

3.2.2 Protein-ligand Interfacial Descriptors

ENTess descriptors

The ENTess chemical geometrical descriptors are obtained by combining Pauling electronegativity (EN) as atomic property and Delaunay Tessellation (Tess) to characterize the protein-ligand interface as follows. When applied to protein-ligand

complexes represented at the atomic resolution level, Delaunay tessellation partitions the protein ligand interface into an aggregate of space-filling, irregular tetrahedra, with both protein and ligand atoms as vertices (see also **Figure 3.3**). Each Delaunay quadruplet is characterized by its unique four-atom composition, which defines the descriptor type (certainly, the same four-body compositions may occur in different, or even the same, protein/ligand interfaces). Furthermore, for each quadruplet we calculate the sum of EN values of the composing atom-vertices, which produces the descriptor value.

PL/MCT descriptors

The PL/MCT descriptors are methodologically similar to ENTess descriptors but are theoretically more rigorous.[154] This is because these new descriptors employ pairwise atomic potentials for the protein-ligand (PL) complexes based on maximal charge transfer (MCT) derived from conceptual Density Function Theory (DFT)[152] in place of Pauling EN, called here PL/MCT. Compared to Pauling EN empirical scales, the conceptual DFT can evaluate chemical properties of different chemical species (e.g., atoms, functional groups, and molecules) systematically.[154] The PL/MCT value is calculated from the following equation (**Equation 3.1**):

$$\mathrm{PL/MCT}_m = \sum_{k=1}^{n} \sum_{p}^{1\sim3} \sum_{l}^{1\sim3} (\mathrm{MCT}_p * \mathrm{MCT}_l / d_{pl})_k \qquad (3.1)$$

where $\mathrm{PL/MCT}_m$ is the potential of the *m*-th tetrahedron type (i.e. individual descriptor type); *n* is the number of occurrences of this tetrahedron type in a given pose; *p* is the vertex index of a protein atom, *l* is the vertex index of a ligand atom, and d_{pl} is the distance between a pair of protein and ligand atoms found in the same Delaunay tetrahedron. (Note that Delaunay tetrahedra at the protein-ligand interface can be classified based on the relative content of protein and ligand atoms, i.e., three protein and

one ligand atoms, two from each, or one protein and three ligand atoms; this explains the tetrahedral type counts in the second and third sum in Eq. 3.1).

The MCT characterizes the maximal electron flow between the donor and acceptor atoms at the protein-ligand interface. It is derived from the conceptual DFT,[152, 155] which provides a theoretical basis for calculating the PL/MCT descriptors. The MCT is calculated as follows, assuming that the total energy of the system is perturbed by the charge transfer up to the second order:

$$\Delta E = \mu \Delta N + 1/2 \eta \Delta N^2 \quad (3.2)$$

where ΔE and ΔN represent energy change and charge transfer, respectively. When the total energy is minimized with respect to the charge transfer, $d\Delta E/d\Delta N = 0$, we have

$$\Delta N_{max} = - \mu / \eta \equiv MCT \quad (3.3)$$

where μ and η are the chemical potential (negative of electronegativity) and the chemical hardness respectively, defined by $\mu = (\partial E/\partial N)_v$ and $\eta = (\partial^2 E/\partial^2 N)_v$ with ν representing the external potential formed by the framework of atomic nuclei.

Occurrence Descriptors

The occurrences of tetrahedral descriptor types at the interface, which are fundamental for calculation of ENTess and PL/MCT descriptors, can be also independently employed as descriptor values in QSBAR modeling.

Combination of ENTess and PL/MCT descriptors

Since the Pauling EN and MCT values represent chemical properties based on distinctive theories, it is sensible to test the modeling performance using the combined descriptor set. The combined descriptor set is constructed by concatenating the ENTess and the PL/MCT descriptor set. We remove the descriptors in the combined descriptor

set which are with low variance (all, or all but one value is constant) and high correlation (if pair-wise square correlation coefficient is greater than 0.99, one of the pair, chosen randomly, is removed). The remaining descriptors are range scaled (0 to 1).

3.2.3 *k*-Nearest Neighbors (*k*-NN) QSBAR Modeling

Phase I

The five-fold external validation technique is applied in the QSBAR modeling. The 665 QSBAR modeling data set is divided, by random selection, into five nearly equal subsets (133 complexes). By turns, one out of five subsets is used solely as the external validation set and the remainder (4/5, 532 complexes) is used for training by the *k*NN algorithm with different descriptor sets. Moreover, the previously excluded complexes of the refined set (635 complexes), whose protein binding sites are dissimilar to the ones in the core set, are retained as an additional external validation set (**Figure 3.4A**).

In addition, the CSAR data sets are utilized as another external validation set for those models that perform best in five-fold external validation. Since many complexes in CSAR data sets have been included in the PDBbind modeling set, their predictions are removed when calculating the prediction statistics.

Phase II

In the second stage of model building, we build models using either Set1 or Set2 data set with the descriptor set, which performs best in the five-fold external validation of Phase I. Then we use the respective models to predict either Set2 or Set1 data set. N-fold external validation technique is also applied except that instead of five-fold validation, ten-fold (Set1) and nine-fold (Set2) are used due to the smaller size of data sets (**Figure 3.4B**).

Phase III

Finally, the PDBbind modeling set is combined with either Set1 or Set2 data set to have a composite modeling set. The resulting models validated by the five-fold external validation technique are employed to predict Set2 or Set1 data set respectively (**Figure 3.4C**).

3.2.4 *k*-NN Modeling Algorithm

Initially, a subset of *nvar* (number of selected variables) descriptors is selected randomly. The model developed with this set of descriptors is validated by leave-one-out (LOO) cross-validation, where each compound is eliminated from the training set and its biological activity is predicted as the weighted average activity of its *k* (*k*= 1 to 9) nearest neighbors in the subspace of *nvar* descriptors (**Equation 3.4**). The weights of neighbors, w_i, decrease with distance, thus closer neighbors contribute to the calculated activity more:

$$y_{pred} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}; \qquad w_i = \exp(-d_i)$$

(3.4)

Here y_{pred} is predicted activity; d_i , w_i and y_i are, respectively: Euclidean distance, weight and actual activity for the nearest neighbor *i*. A genetic algorithm was used to optimize the variable selection (with population size of 500 solutions of *nvar* size from 5 to 50 descriptors).

3.2.5 Validation of QSBAR Models

As emphasized in the previous study, training-set-only modeling is insufficient to achieve models with validated predictive power. Thus, prior to *k*NN, the modeling set (532 complexes) is further subdivided into 20 training/test subsets using the sphere exclusion algorithm[20], maximizing the diversity of both training and test sets. The *k*NN QSBAR models are developed solely based on these training sets and the resulting models are validated through predicting the binding affinity of complexes in the respective test sets. The statistical significance of QSBAR models is characterized by the following parameters: a) LOO cross-validated q^2; b) square of the correlation coefficient R (R^2) between the predicted and observed activities; c) coefficients of determination (predicted *vs.* observed activities R_0^2, and observed *vs.* predicted activities R'^2_0; d) slopes k and k' of regression lines (predicted *vs.* observed activities, and observed *vs.* predicted activities) through the origin.

The detailed discussion of these parameters has been provided in previous studies.[20, 126]

Individual models are considered to have acceptable predictive power if

$$q^2 \geq 0.5 \text{ and } R^2 \geq 0.6$$

otherwise they are discarded. The ensemble of all models that pass the above criteria is then used for consensus prediction of compounds in an external validation set.

Besides, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values are also used in evaluating the prediction results of five-fold external cross validation.

3.2.6 Applicability Domain

Because *k*NN models interpolate activities from the nearest neighbor compounds in the relevant training sets, a similarity threshold (i.e., model applicability domain) should be introduced to avoid making predictions for compounds that differ substantially from the training set molecules. The similarity threshold is defined as follows:

$$D_T = \bar{y} + Z\sigma \quad (3.5)$$

Here, $\bar{y}$ is the average over Euclidean distances to *k* nearest neighbors of all compounds in the training set (where the value of k is the same as in predictive *k*NN QSBAR models), σ is the corresponding standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. Typically, we set Z to 0.5, which places the boundary for deciding whether a compound is within or outside of the applicability domain at one-half of the standard deviation. It is important to notice that increasing the value of Z would increase the number of compounds in the external set that are considered within the applicability domain but could decrease the accuracy of prediction due to inclusion of dissimilar nearest neighbors. Previous studies have demonstrated that model applicability domain is important in model building and should be universally applied during the prediction.[18]

Moreover, we further introduce another similarity threshold to avoid making predictions for compounds that differ substantially from the modeling set compounds (i.e., global applicability domain). The definition of this similarity threshold is analogous to the one of model applicability domain except the average Euclidean distance ($\bar{y}$) and standard deviation (σ) are calculated by using *one* nearest neighbor of each compound within the modeling set in the entire descriptor space.

3.2.7 Stochastic Proximity Embedding

The stochastic proximity embedding (SPE) algorithm intends to embed high-dimensional data points into a low-dimensional space yet preserves the geodesic distances between the embedded data.[156, 157] The classical algorithm for this task is principle component analysis (PCA). However, PCA can fail to provide accurate projection if its first few components do not cover enough of data variation, which is common when handling data sets with diverse chemical structures.

For the SPE calculation, initial 2D coordinates are randomly assigned to each data point and then are refined by iteratively selecting pairs of data points and adjusting their coordinates based on their respective proximities in the original descriptor space. The calculation is carried out using the entire set of descriptors computed for the modeling set (e.g., Set1) and its respective external validation set (e.g., Set2). The SPE calculations are conducted by in-house Matlab (version 7.7.0)[158] scripts.

3.3 Results and Discussions

3.3.1 Assessment of Protein-ligand Interfacial Descriptors Performance

The statistics of five-fold external validation results of PDBbind data set are presented in **Table 3.2**. These predictions are made by qualified models ($q^2 \geq 0.5$ and $R^2 \geq 0.6$) based on different sets of descriptors. Generally, the predictions of five-fold external sets using models built by ENTess or PL/MCT descriptors show only marginal improvement when compared with occurrence descriptors. In contrast, significant improvement of five-fold external predictions in all statistical metrics is observed using the models built with the combined descriptor set (ENTess + PL/MCT), demonstrating

the synergetic effect of ENTess and PL/MCT descriptors in model building. Then we apply all the models generated from five splits by different descriptor sets to predict binding affinity of complexes in the additional external validation set, whose protein binding sites are dissimilar to the core set. The prediction accuracy of this external set drops drastically yet the predictions from models built with the combined descriptor set are still slightly better than for other descriptor sets (**Table 3.3**). This result suggests that it should be possible to improve external prediction accuracy by removing complexes whose binding sites are dissimilar to the PDBbind modeling set. Since the models built by the combined descriptor set show better performances in predicting both five-fold external validation sets and the additional validation set, the combined descriptor set is employed in further Phase II and Phase III model building.

3.3.2 Model Validation Using CSAR Data Sets

The phase I models built with the combined descriptor set are also used to predict the CSAR data sets. However, since part of the CSAR data sets originates from the PDBbind database, many predictions cannot be considered as rigorous (i.e., the overlaps between the modeling set and the validation set) and are removed when calculating the prediction statistics. The results are shown in **Table 3.5**. The R^2 for Set1 is somewhat better than for Set2 (0.48 *versus* 0.42) despite the fact that Set1 data set is more diverse in nature. Nevertheless, both sets' predictions are much worse than the results of five-fold external validation using PDBbind modeling set.

Since the CSAR exercise requires participants to submit valid predictions of *all* CSAR complexes for analysis, the Set1 and Set2 data set are re-predicted by models built from either Set2 or Set1 data set respectively (phase II in **Figure 3.5**). The results of

external n-fold cross validation from CSAR data set modeling are reported in **Figure 3.5**A and **Table 3.4**. The average R^2 is 0.45 for Set1 data set modeling and 0.53 for Set2 data set modeling in external n-fold cross validation, in agreement with the fact that the number of activity outliers (usually called "activity cliffs") in Set1 data set is larger than in Set2 data set (**Figure 3.6**). Then the validated Set1 or Set2 models are applied to predict Set2 and Set1 data set respectively (**Table 3.5**). Interestingly, the R^2 of Set2 prediction using Set1 models is significantly better than the one using PDBbind models (0.51 *versus* 0.42) even though there are more complexes employed in PDBbind model building. On the other hand, the R^2 of Set1 data set prediction using Set2 models is much worse than the one using PDBbind models (0.40 *versus* 0.48).

Neither PDBbind models nor CSAR models can have desirable prediction reliability for the CSAR data sets. Therefore, we manage to build the QSBAR models using the data set by combining the PDBbind data set with either Set1 or Set2 data set (phase III). The external cross validation results are reported in **Figure 3.5B** and **Table 3.4**. The average R^2 is 0.56 for PDBbind plus Set1 data set modeling and 0.59 for PDBbind plus Set2 data set modeling in external five-fold cross validation. Then the validated models built on PDBbind plus Set1 and on PDBbind plus Set2 are applied to predict Set2 and Set1 respectively. The prediction of both Set1 and Set2 data sets shows improved statistics compared to previous results using PDBbind Phase I models or CSAR Phase II models (**Table 3.5**). The prediction accuracy (R^2) can be as high as 0.50 for Set1 (*versus* 0.48 by PDBbind models and 0.40 by Set2 models) and 0.53 for Set2 (*versus* 0.42 by PDBbind models and 0.51 by Set1 models). We suspect the improvement might be due to including into the modeling set more complexes that are structurally similar to the

complexes in the external set as well as due to the decreased number of activity outliers in the modeling set (*vide infra*).

3.3.3 Analysis of Nearest Neighbor Distribution of CSAR Data Sets

We compare nearest neighbor (NN) distribution within Set1 (Set2) as an external set to the distribution of NNs between external and modeling sets, where modeling set can be Set2 (Set1) or PDBbind + Set2 (Set1). The results are shown in **Figure 3.6**, where we plot the pairwise nearest neighbor distances, adjusted based on the number of descriptors in the modeling set, *versus* pairwise binding affinity difference. The mean and standard deviation of nearest neighbor distances are calculated (mean.NN.dist and std.NN.dist.) and reported in **Table 3.6**.

We have tried to analyze if the improvement of external prediction accuracy using models built from PDBbind plus Set2 (Set1) data set is due to the decreased number of activity outliers. We define activity outliers as the data points whose distance to the nearest neighbors is small while binding affinity difference is large (for example, see the green shaded area in **Figure 3.6**). An overview of **Figure 3.6** shows that the nearest neighbor distribution of Set1 external validation set from the PDBbind plus Set2 modeling set is more compact in comparison with the one from the Set2 modeling set, tending to include more structurally similar complexes and have fewer activity outliers. This agrees with the results that the external prediction R^2 of Set1 data set by models built from PDBbind plus Set2 is better than by models built from Set2 alone (0.50 *versus* 0.40). Regarding the nearest neighbor distribution of Set2 external validation set from the PDBbind plus Set1 modeling set, and from the Set1 modeling set, the difference in

number of activity outliers from two distributions is less obvious, which is in line with the smaller difference between external prediction accuracy (R^2: 0.53 *versus* 0.51)

3.3.4 The Effect of Applicability Domain

The Applicability Domain (AD) defines the area of the descriptor space in which QSBAR models can predict the binding affinity of complexes more reliably. If a protein-ligand complex has interfacial chemical geometry that is "too dissimilar" to that of all complexes in the modeling set (i.e., greater than the predefined similarity threshold, Z-cutoff=0.5), we assume that we cannot predict its activity reliably. The model AD is universally applied, but the application of global AD is switched on and off during the prediction in order to study its impact.

In Phase I modeling, after applying the global AD, no matter which descriptor set is used for constructing models, we observe consistent improvement of prediction accuracy in five-fold external validation (**Table 3.2**). However, the number of complexes that can be predicted (i.e., coverage) decreases after the application of global AD. The best average R^2 across five folds is 0.68 (with global AD) in comparison with 0.63 (w/o global AD) by using models built with the combined descriptor set. Similarly, the prediction of those complexes with pockets dissimilar to the core set improves (e.g., from 0.28 to 0.40) after applying the global AD (**Table 3.3**).

When predicting the CSAR data sets, we also apply the global AD which is defined based on PDBbind data set, Set1/Set2 data set, or PDBbind plus Set1/Set2 data set (cf. **Table 3.5**). A tangible improvement of R^2 is observed for prediction of both sets using models built from either PDBbind data set (Set1: 0.48 to 0.53 and Set2: 0.42 to 0.47) or PDBbind plus Set1/Set2 data set (Set1: 0.50 to 0.56 and Set2: 0.53 to 0.58). The

overall prediction accuracy is comparable (despite the lower coverage rate) to the best in CSAR exercise reported in July 2010. However, the global AD fails to improve the results when the predictions are made by using models built from either Set1 or Set2 data set.

Since the global AD is defined using the average and standard deviation of nearest neighbor distances based on the entire descriptor space of the modeling set and is then applied to the external set to exclude complexes which are dissimilar to the modeling set, we would like to examine the relative distribution of external set and modeling set data points. Initially we applied the PCA analysis since it is the most popular approach to visualize relative distribution of data points in 3D space. However, only less than 30% of variance within the data set can be explained by the first three principle components (PC). Thus, PCA's visualization is not representative. Instead, we employed the SPE method, which can preserve the intrinsic relationships of high-dimensional data.

As shown in the SPE plots (**Figure 3.7**), the data distribution of Set1 external set is much sparser in the Set2 descriptor space compared with the one in the PDBbind descriptor space or in the PDBbind plus Set2 descriptor space. Similarly, the data distribution of Set2 modeling set is sparser than the one of PDBbind or PDBbind plus Set2 modeling set (**Figure 3.7A**). Incidentally, the R^2 of Set1 prediction using models built from Set2 drops from 0.4 to 0.3 after applying global AD. At the same time, there is significant R^2 improvement of Set1 prediction, after applying global AD, using models built either from PDBbind data set (R^2 from 0.48 to 0.53) or from PDBbind plus Set2 (from 0.50 to 0.56). On the other hand, data distribution of Set2 in the Set1 descriptor

space is relatively sparse and applying global AD did not help improve the Set2 prediction accuracy by models built on Set1 data set (from 0.51 to 0.50). However, a significant R^2 improvement is seen after applying global AD when models built from PDBbind dataset (from 0.42 to 0.47) or from PDBbind plus Set1 (from 0.53 to 0.58) are used in prediction. This analysis might indicate the limitation of the current global AD definition as it seems to be sensitive to changes in data set distribution.

3.4 Conclusions

We have modified previous ENTess descriptors by incorporating theoretically more rigorous values (i.e., conceptual DFT atomic properties) as well as protein-ligand pairwise distances within tetrahedra into descriptor generation. We named the new descriptors as PL/MCT descriptors. Employing models built by PL/MCT descriptors in combination with ENTess descriptors, the prediction accuracy in five-fold external validation of PDBbind data set is much better than using models built by any single descriptor set. Furthermore, we applied this combined descriptor set to construct models to predict CSAR data sets. When predicting the CSAR data sets (Set1 or Set2), we got better prediction accuracy by using models built by data set including both PDBbind data set and CSAR data set (Set2 or Set1) than by using models built by either PDBbind data set or CSAR data set (Set1 or Set2) alone, indicating the model quality and applicability are improved. This improvement seems to be due to the inclusion of more complexes structurally similar to the prediction set as well as due to the decreased number of activity outliers. Moreover, although applying global applicability domain decreases the prediction coverage, it also can help to significantly improve the prediction accuracy in some cases. The overall R^2 of external sets in CSAR

exercise can be as high as 0.57, which is comparable to the best in the CSAR exercise (R^2 = 0.58, July 2010). However, we also demonstrate that applying global applicability domain (as it is defined in Methods) does not help or even deteriorate the prediction accuracy when the data distribution of external sets and/or modeling set is very sparse.

Figures for Chapter 3

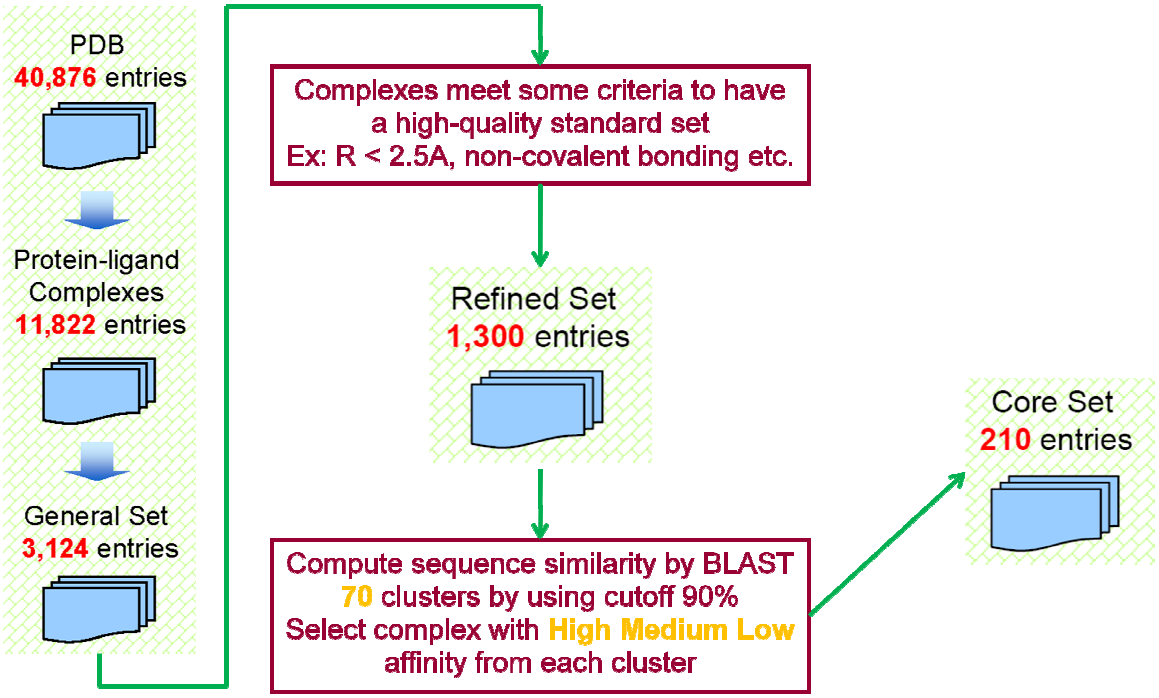


Figure 3.1: A brief introduction to the PDBbind v. 2007.

(The graph was modified from
http://sw16.im.med.umich.edu/databases/pdbbind/pdfs/pdbbind_2007_intro.pdf)

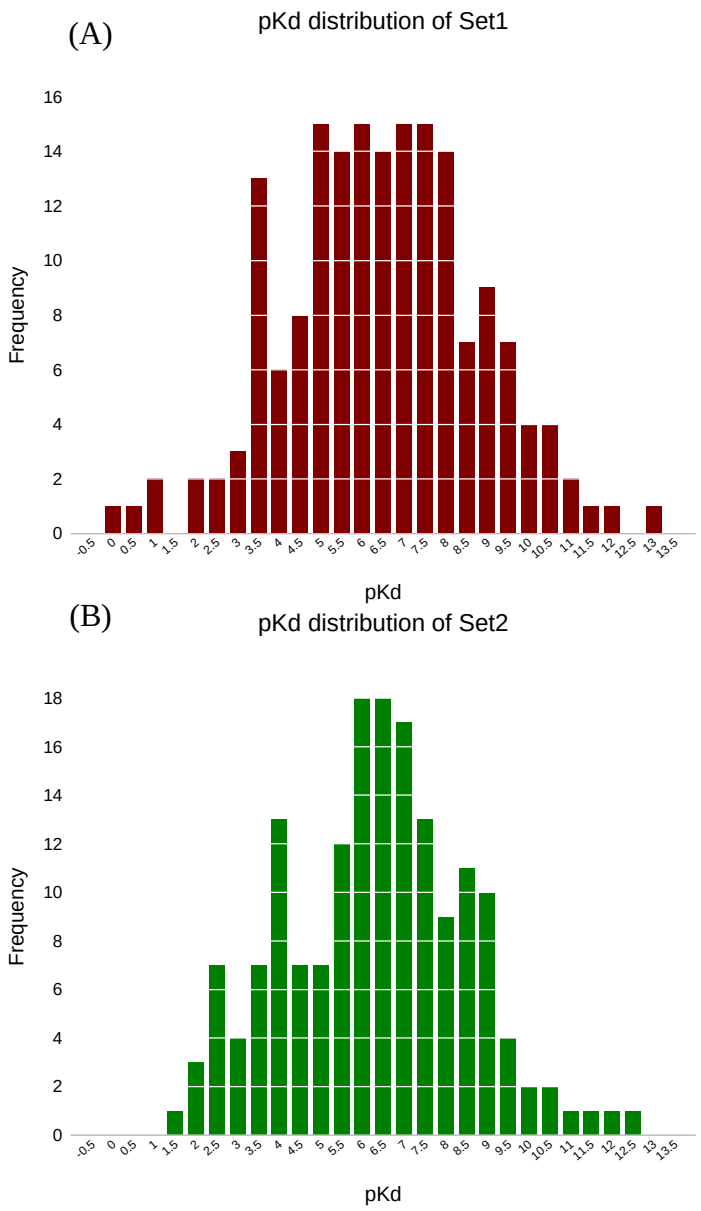


Figure 3.2: The pK_d distribution of CSAR data sets (A. Set1; B. Set2).

The x-axis is the pK_d value binned by 0.5 log value and the y-axis is the frequency of data points in the corresponding bin.

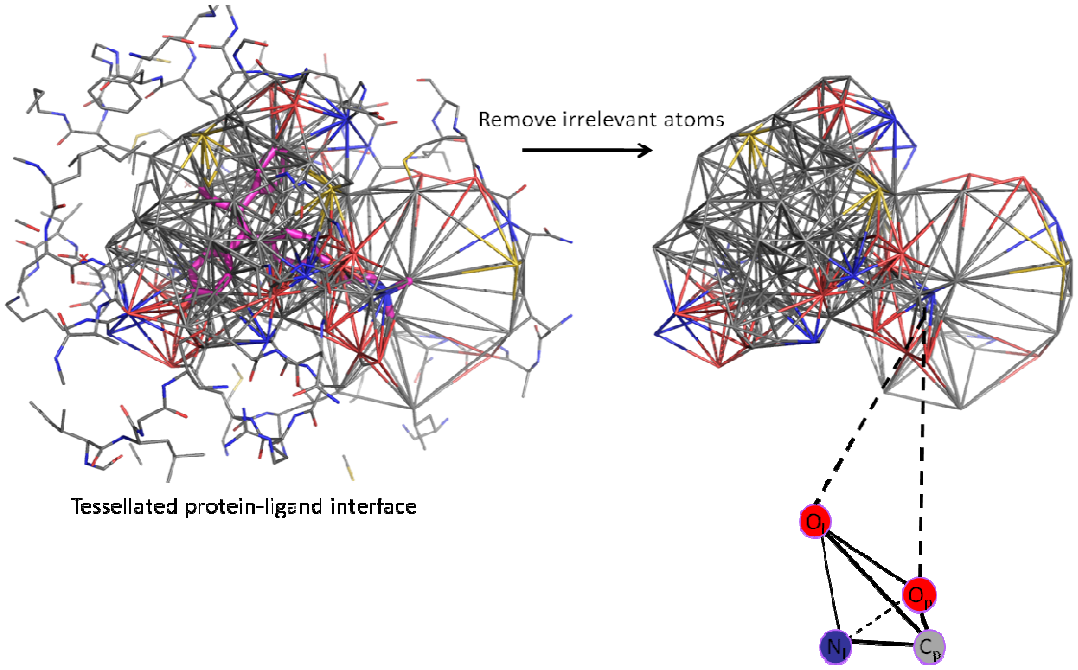
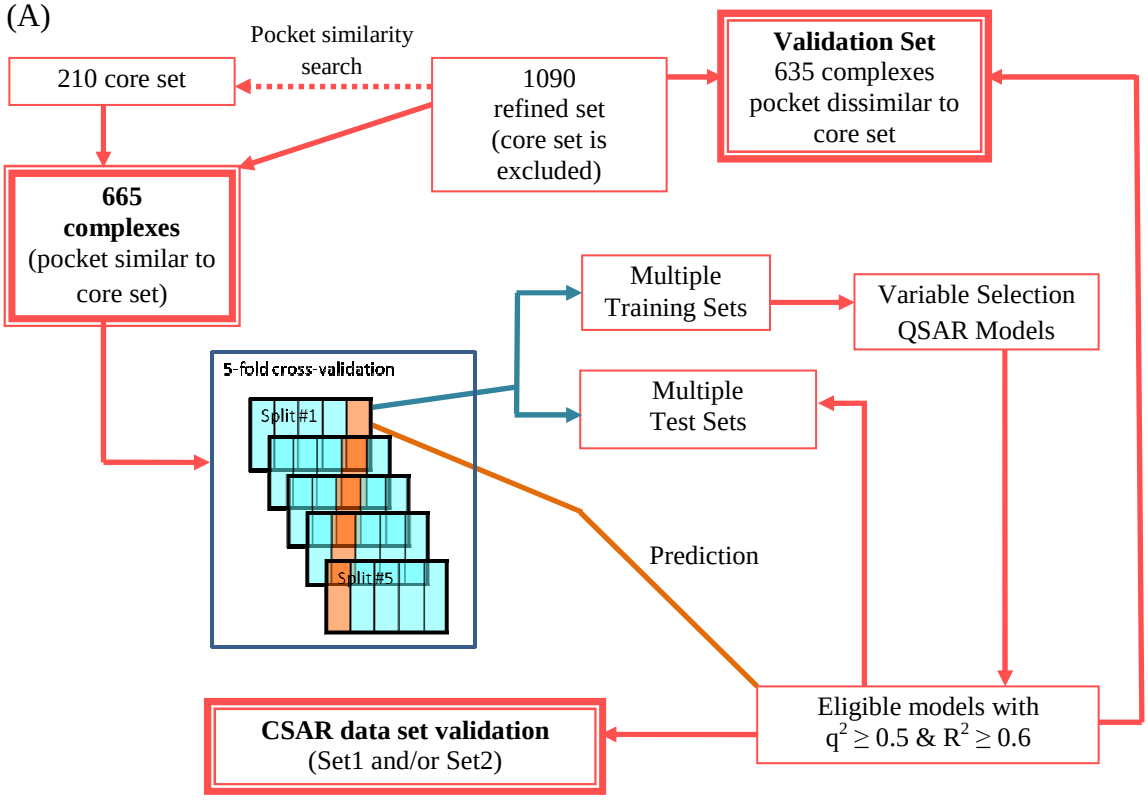


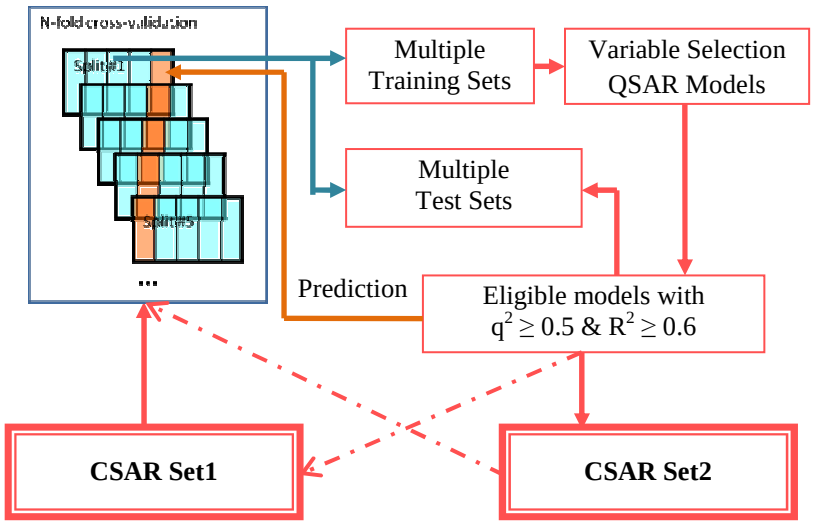
Figure 3.3: Illustration of the method to derive PL/MCT descriptors using the tesselated protein-ligand complex (3ERT, the ER/antagonists benchmarking dataset).

The atom types for protein and ligand are treated differently. For instance, for the tetrahedron at the left corner, C_p and O_p are carbon and oxygen atoms from the protein while O_l and N_l are oxygen and nitrogen atoms from the ligand.

(A)


Pocket similarity search
210 core set
1090 refined set (core set is excluded)
Validation Set 635 complexes pocket dissimilar to core set
665 complexes (pocket similar to core set)
Multiple Training Sets
Variable Selection QSAR Models
5-fold cross-validation
Split#1
Split#5
Multiple Test Sets
Prediction
Eligible models with $q^2 \geq 0.5$ & $R^2 \geq 0.6$
CSAR data set validation (Set1 and/or Set2)

(B)


N-fold cross-validation
Split#1
Split#5
...
Multiple Training Sets
Variable Selection QSAR Models
Multiple Test Sets
Prediction
Eligible models with $q^2 \geq 0.5$ & $R^2 \geq 0.6$
CSAR Set1
CSAR Set2

(C)

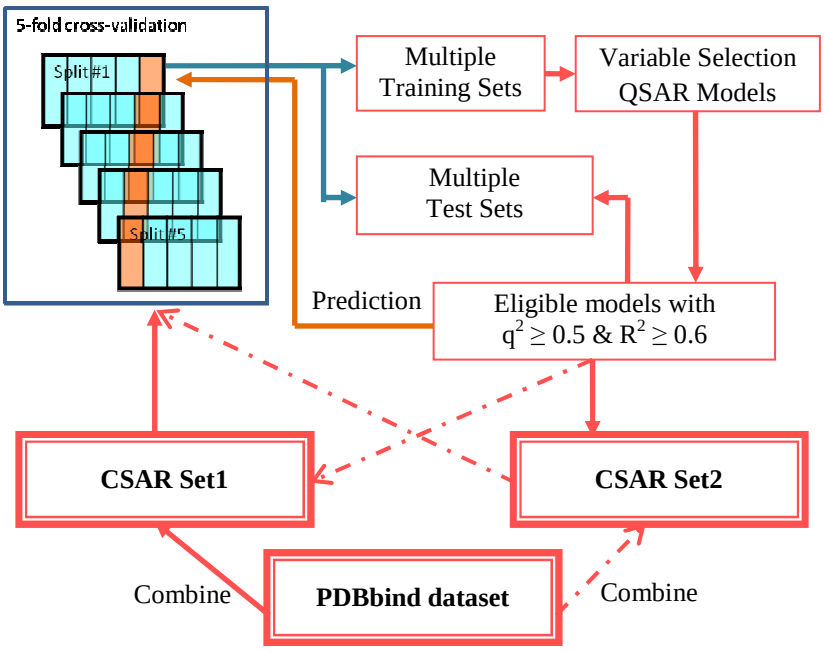
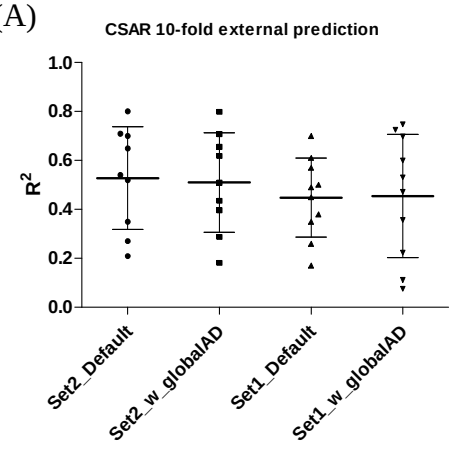
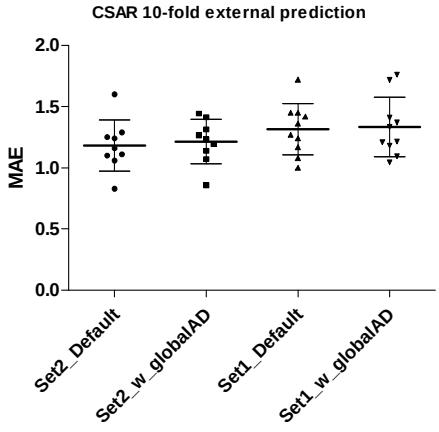
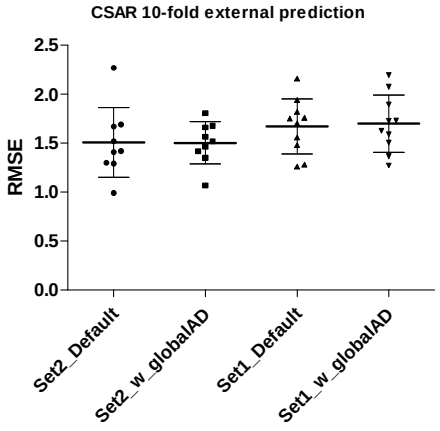


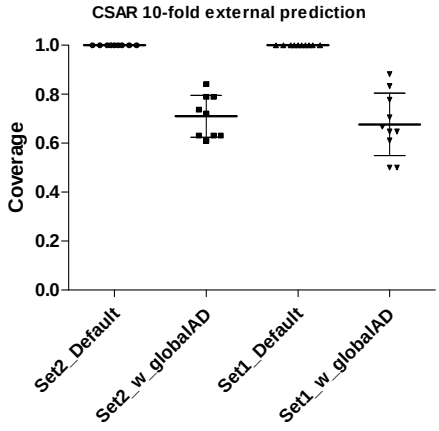
Figure 3.4: The workflow of model building and validation using A) PDBbind data set; B) Set1 (solid line) or Set2 (dash-dotted line); C) PDBbind plus Set1 (solid line) or PDBbind plus Set2 (dash-dotted line).

(A)


CSAR 10-fold external prediction
R2
1.0
0.8
0.6
0.4
0.2
0.0
Set2_Default
Set2_w_globalAD
Set1_Default
Set1_w_globalAD


CSAR 10-fold external prediction
MAE
2.0
1.5
1.0
0.5
0.0
Set2_Default
Set2_w_globalAD
Set1_Default
Set1_w_globalAD


CSAR 10-fold external prediction
RMSE
2.5
2.0
1.5
1.0
0.5
0.0
Set2_Default
Set2_w_globalAD
Set1_Default
Set1_w_globalAD


CSAR 10-fold external prediction
Coverage
1.0
0.8
0.6
0.4
0.2
0.0
Set2_Default
Set2_w_globalAD
Set1_Default
Set1_w_globalAD

(B)

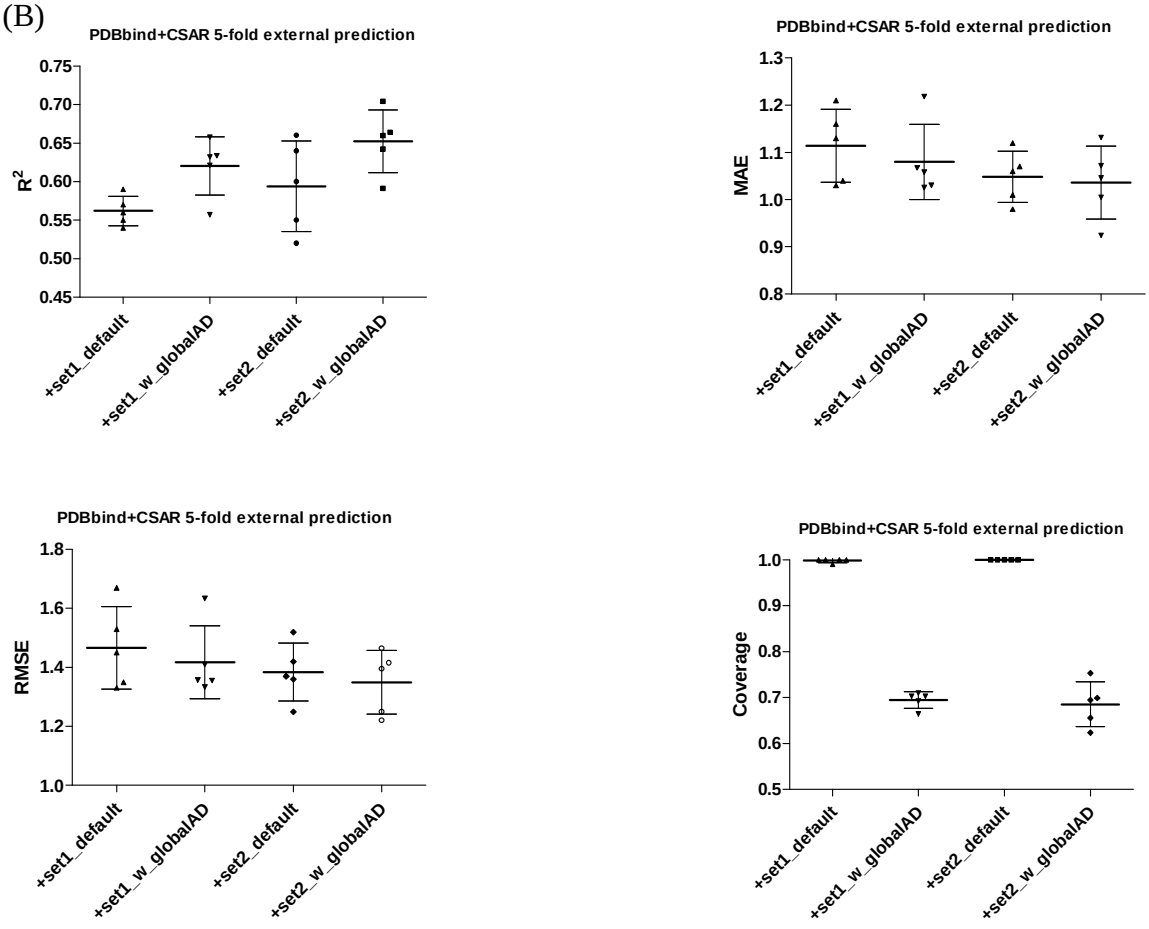


Figure 3.5: The statistics (R^2, MAE, coverage, and RMSE; clockwise) of external n-fold validation sets using models built with A) Set1 (or Set2); B) PDBbind plus Set1 (or PDBbind plus Set2).

The mean and $\pm\mathbf{1}$ standard deviation of data points are shown as horizontal lines on each plot. The improvement of average prediction accuracy is negligible when applying global AD to models built with Set1 (or Set2) alone.

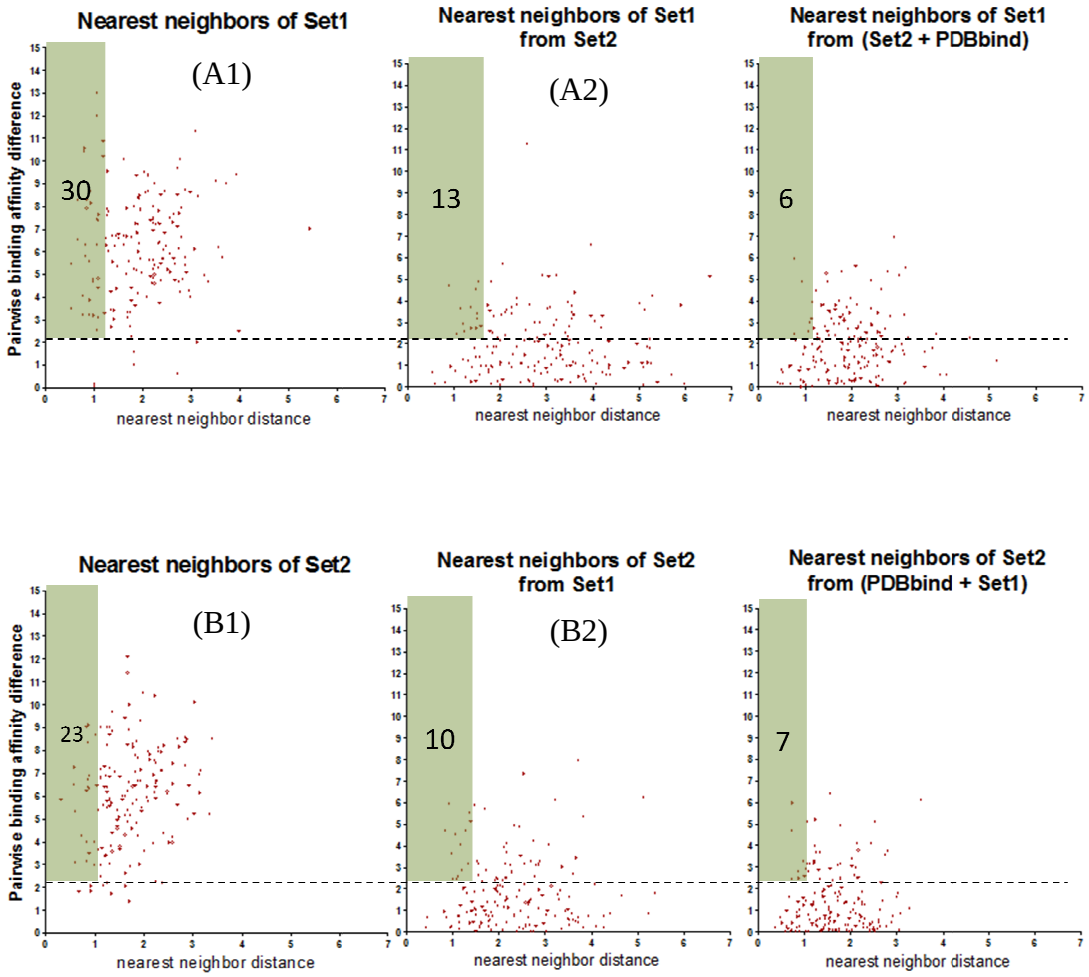
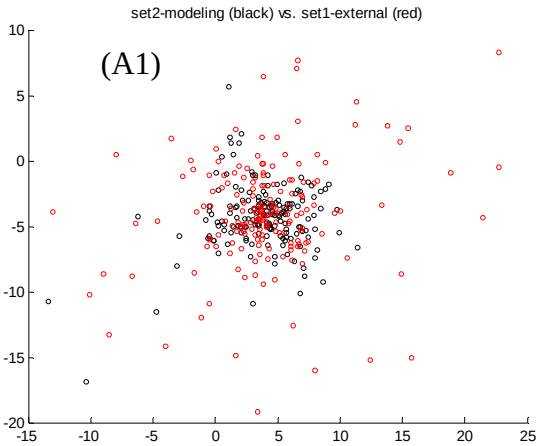
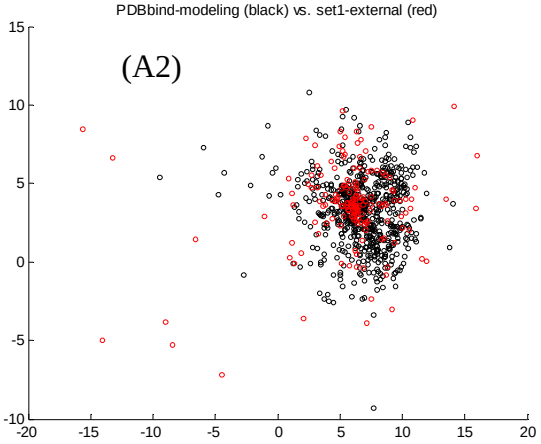
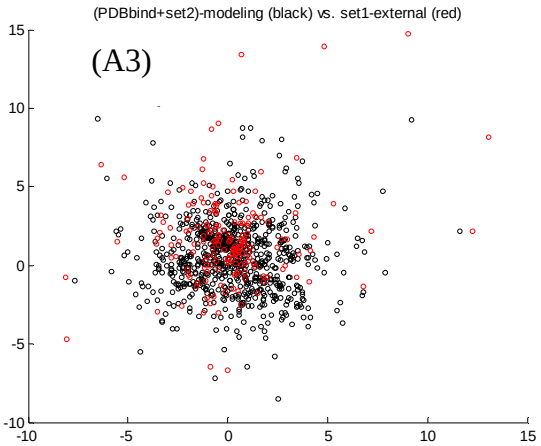


Figure 3.6: Nearest neighbor distribution of Set1 as external set: A1) within itself; A2) based on neighbors taken from Set2 modeling set; A3) based on neighbors taken from PDBbind + Set2 modeling set. Likewise, nearest neighbor distribution of Set2 external set: B1) within itself; B2) based on neighbors from Set1 modeling set; A3) based on neighbors from PDBbind + Set1 modeling set.

The x-axis is adjusted nearest neighbor distance and the y-axis is the binding affinity difference between each complex in the external set and its nearest neighbor. The dashed line marks the standard deviation of binding affinities in the external set (cf. **Table 3.1**); the green shade covers the region of "activity cliff"; the number in the green shade is the number of activity outliers (defined as: y < mean.NN.dist. – std.NN.dist. and x > std. binding affinity of complexes within the external set).


set2-modeling (black) vs. set1-external (red)
(A1)


PDBbind-modeling (black) vs. set1-external (red)
(A2)


(PDBbind+set2)-modeling (black) vs. set1-external (red)
(A3)

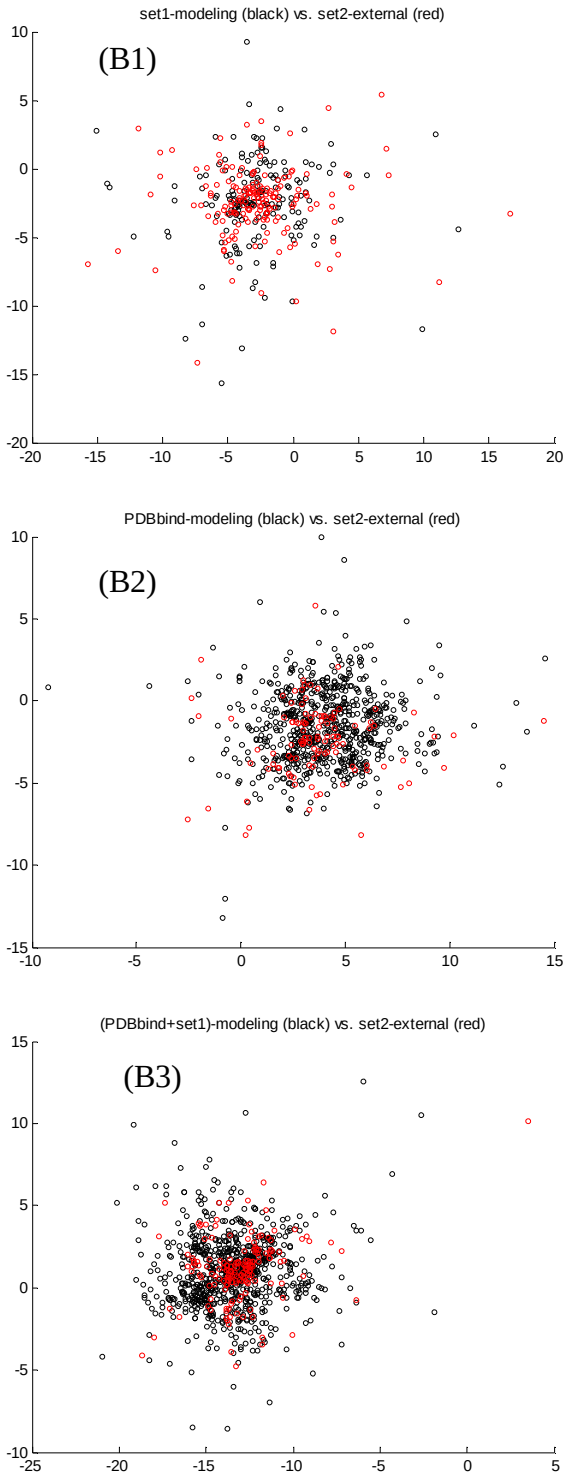


Figure 3.7: The 2D SPE plots. The black dots are data points of the external set and the red dots are data points of the modeling set.

Set1 is the external set in A plots; Set2 is the external set in B plots. Plots 1-3 represent the data distribution in the descriptor space of different modeling sets: 1, Set2 (Set1); 2, PDBbind set; 3, PDBbind plus Set2 (Set1). The absolute coordinates of data points are generated randomly then iteratively optimized. Only relative positions of data points are meaningful.

Tables for Chapter 3

Table 3.1: The discriminant analysis of data sets based on protein-ligand binding pK_d values and protein sequences

	data set / parameter	Set1	Set2	PDBbind	Old ENTess
pK_d values	Count	176	169	665	264
	Mean	6.23	6.10	6.69	6.42
	Median	6.25	6.24	6.77	6.57
	Standard deviation	2.31	2.17	2.22	2.39
	Range/Lowest/Highest	13.15/-0.15/13	10.7/1.4/12.1	12.6/1.36/13.96	12.48/1.48/13.96
sequence	# of families/ # of singletons (90% sequence similarity)	121/80	107/68	101/26	83/43

Table 3.2: The statistics (R^2, coverage, MAE, and RMSE) of five-fold external validation sets using models built with PDBbind data set using occurrence, ENTess, PL/MCT, or combined descriptor set (ENTess + PL/MCT).

R^2:					
Descriptor \ Fold	#1	#2	#3	#4	#5
Occurrence	0.55/0.63*	0.57/0.59	0.56/0.60	0.57/0.61	0.55/0.57
ENTess (1)	0.56/0.66	0.59/0.61	0.54/0.62	0.63/0.69	0.53/0.56
PL/MCT (2)	0.54/0.62	0.58/0.58	0.56/0.59	0.62/0.68	0.54/0.55
(1) + (2)	**0.63/0.70**	**0.64/0.67**	**0.62/0.69**	**0.68/0.72**	**0.56/0.61**
coverage					
Descriptor \ Fold	#1	#2	#3	#4	#5
Occurrence	1/0.64	0.99/0.72	0.99/0.78	0.99/0.71	0.95/0.73
ENTess (1)	1/0.64	0.99/0.73	0.99/0.79	0.99/0.70	0.98/0.71
PL/MCT (2)	1/0.66	0.99/0.74	0.98/0.77	0.99/0.70	0.99/0.72
(1) + (2)	1/0.64	1/0.67	1/0.75	1/0.73	1/0.68
MAE					
Fold \ Descriptor	#1	#2	#3	#4	#5
Occurrence	1.09/0.98	1.09/1.08	1.21/1.22	0.98/0.99	1.14/1.11
ENTess (1)	1.12/0.98	1.08/1.07	1.26/1.20	0.92/0.89	1.19/1.10
PL/MCT (2)	1.11/1.00	1.09/1.04	1.22/1.21	0.93/0.89	1.19/1.16
(1) + (2)	**0.99/1**	**1.04/1.02**	**1.15/1.02**	**0.84/0.79**	**1.12/1.17**
RMSE					
Fold \ Descriptor	#1	#2	#3	#4	#5
Occurrence	1.44/1.35	1.47/1.51	1.57/1.55	1.29/1.28	1.51/1.44
ENTess (1)	1.41/1.26	1.42/1.46	1.61/1.51	1.18/1.15	1.56/1.45
PL/MCT (2)	1.46/1.36	1.45/1.44	1.57/1.59	1.20/1.16	1.53/1.49
(1) + (2)	**1.30/1.26**	**1.36/1.39**	**1.46/1.28**	**1.12/1.08**	**1.50/1.62**

*default/with global AD, Z = 0.5; bold type: the best statistics in folds and global AD helps improve the results

Table 3.3: The statistics (R^2, coverage, MAE, and RMSE) of external validation set (complexes which have pockets dissimilar to the core set) using models built with PDBbind data set using occurrence, ENTess, PL/MCT, or combined descriptor set (ENTess + PL/MCT)

Parameter / Descriptor	R^2	Coverage	MAE	RMSE
Occurrence	0.24/0.32*	1/0.30	1.39/1.31	1.84/1.78
ENTess (1)	0.24/0.35	1/0.29	1.37/1.29	1.83/1.71
PL/MCT (2)	0.26/0.34	1/0.32	1.35/1.3	1.81/1.76
(1) + (2)	**0.28/0.40**	**1/0.39**	**1.34/1.23**	**1.77/1.57**

*default/ with AD, Z = 0.5; bold type: the best statistics

Table 3.4: The statistics (R^2, MAE, coverage, RMSE, and coverage) of external n-fold validation sets using models built from Set1, Set2, PDBbind plus Set1, or PDBbind plus Set2.

"Set1" data set modeling										
Fold / Parameter	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
R^2	0.38/0.47	0.17/0.22	0.50/0.53	0.49/0.74	0.45/0.07	0.57/0.60	0.26/0.36	0.70/0.70	0.61/0.72	0.35/0.11
MAE	1.00/1.04	1.27/1.41	1.72/1.72	1.45/1.18	1.45/1.76	1.24/1.21	1.42/1.33	1.08/1.21	1.17/1.09	1.36/1.3
RMSE	1.28/1.27	1.56/1.73	2.16/2.19	1.94/1.63	1.75/2.08	1.76/1.73	1.82/1.59	1.26/1.36	1.48/1.51	1.70/1.89
Coverage	1.00/0.70	1.00/0.65	1.00/0.88	1.00/0.65	1.00/0.50	1.00/0.83	1.00/0.61	1.00/0.78	1.00/0.67	1.00/0.50
"Set2" data set modeling										
Fold / Parameter	#1	#2	#3	#4	#5	#6	#7	#8	#9	NA
R^2	0.54/0.51	0.35/0.40	0.65/0.65	0.70/0.71	0.80/0.80	0.21/0.18	0.52/0.43	0.71/0.62	0.27/0.29	NA
MAE	1.10/1.14	1.16/1.31	1.24/1.41	0.83/0.86	1.06/1.19	1.6/1.07	1.11/1.23	1.25/1.27	1.29/1.44	NA
RMSE	1.30/1.35	1.42/1.56	1.52/1.66	0.99/1.07	1.29/1.42	2.27/1.46	1.41/1.52	1.67/1.68	1.69/1.81	NA
Cover	1.00/0	1.00/0	1.00/0	1.00/0	1.00/0	1.00/0	1.00/0	1.00/0	1.00/0	NA

age	.72	.61	.79	.63	.79	.63	.84	.74	.63

*default/ with global AD, Z = 0.5

"PDBbind plus Set1" data set modeling					
Parameter \ Fold	#1	#2	#3	#4	#5
R^2	0.54/0.56*	0.59/0.66	0.55/0.62	0.57/0.63	0.56/0.63
MAE	1.03/1.02	1.13/1.07	1.21/1.22	1.04/1.03	1.16/1.06
RMSE	1.35/1.35	1.45/1.41	1.67/1.63	1.33/1.33	1.53/1.36
Coverage	1.00/0.69	1.00/0.70	1.00/0.71	1.00/0.66	0.99/0.70
"PDBbind plus Set2" data set modeling					
Parameter \ Fold	#1	#2	#3	#4	#5
R^2	0.52/0.59	0.66/0.70	0.60/0.66	0.55/0.64	0.64/0.66
MAE	1.07/1.04	0.98/1.00	1.12/1.07	1.01/0.92	1.06/1.13
RMSE	1.42/1.41	1.25/1.25	1.52/1.39	1.36/1.22	1.37/1.46
Coverage	1.00/0.70	1.00/0.75	1.00/0.69	1.00/0.65	1.00/0.62

*default/ with global AD, Z = 0.5

Table 3.5: The statistics (R^2, R_0^2, coverage, MAE, and RMSE) of Set1 and Set2 prediction using models built from Set2 (or Set1), PDBbind data set, and PDBbind plus Set2 (or Set1) with combined descriptor set (ENTess + PL/MCT)

Parameters	R^2	R_0^2	RMSE	MAE	Coverage
Models	Set1 prediction				
Set2	0.40/0.30*	0.40/0.29	1.82/1.93	1.36/1.46	1/0.47
PDBbind	0.48/0.53	0.48/0.53	1.68/1.75	1.32/1.44	1^+/0.49
PDBbind + Set2	**0.50/0.56**	**0.50/0.55**	**1.62/1.67**	**1.26/1.30**	**1/0.49**
Models	Set2 prediction				
Set1	0.51/0.50	0.50/0.50	1.53/1.55	1.18/1.23	**1/0.73**
PDBbind	0.42/0.47	0.41/0.47	1.59/1.54	1.22/1.22	1^+/0.64
PDBbind + Set1	**0.53/0.58**	**0.53/0.57**	**1.49/1.46**	**1.14/1.15**	1/0.67

*default/ with AD, Z = 0.5; bold type: the best statistics and
global AD helps improve the results

$^+$ only 169 out of 176 in Set1 and only 122 out of 169 in Set2 are not overlapped with PDBbind modeling set

Table 3.6: Analysis of nearest neighbors of Set1 (Set2), as external validation set, taken from itself, from Set2 (Set1), or from PDBbind plus Set2 (Set1) and the prediction accuracy of Set1 (Set2) external validation set using models built from Set2 (Set1) modeling set and PDBbind plus Set2 (Set1) modeling set

Nearest neighbors (NN) of Set1 external set			
Data sets→	Set1	Set2	PDBbind+Set2
Mean NN dist.	1.97	3	1.98
Std. NN dist.	0.79	1.34	0.81
R^2 (RMSE)	-	0.40 (1.82)	0.50 (1.62)
Nearest neighbors (NN) of Set2 external set			
Data sets→	Set 2	Set 1	PDBbind+Set1
Mean NN dist.	1.72	2.32	1.68
Std. NN dist.	0.68	0.97	0.66
R^2 (RMSE)	-	0.51 (1.53)	0.53 (1.49)

Std.: standard deviation; dist.: distance

Chapter 4 Cheminformatics Meets Molecular Mechanics: A Combined Application of Knowledge-based Pose Scoring and Physical Force Field-based Hit Scoring Functions Improves the Accuracy of Structure-based Virtual Screening

4.1 Introduction

In recent years, virtual screening (VS) has become an increasingly popular strategy for computer-aided drug design.[28, 159] VS approaches explore available or synthetically feasible chemical databases to identify a relatively small number of high-scoring hits that can be validated experimentally. A successful VS method can be applied to large data sets of compounds, resulting in significant enrichment of true binders among the top ranking hits.

Two types of methodologies are employed in virtual screening: structure-based[160, 161] and ligand-based.[9] Structure-based approaches require knowledge of the 3D structure of the target, and employ docking methods to generate binding poses. Then, scoring functions are used to identify the putative native-like pose(s), for which binding affinities can be predicted. Conversely, most ligand-based VS methods search chemical compound databases to identify molecules that are chemically similar to known active ligands or are predicted to be active against the respective targets. Such methods do not require knowledge of 3D structures of the targets and are computationally efficient.[9] However, ligand-based VS approaches require knowledge of active ligands and have inherently lower potential to identify novel chemical scaffolds than do structure-based methods. Nevertheless, recent studies have shown that ligand-based methods have often outperformed structure-based approaches in terms of VS efficacy.[162]

Rigorous scoring functions are a critical component of structure-based VS approaches. Most scoring functions predict binding affinity using physical force fields that account for intermolecular interactions such as electrostatic, van der Waals and hydrophobic interactions, and hydrogen bonding. Due to the static nature of the underlying molecular models many important effects influencing the binding free energy are often not taken into account; examples include entropy, micro-environment dependent polarization, π-stacking, and solvent effects.

Recent studies have shown that inaccuracy of scoring functions is the major bottleneck of structure-based VS.[47] It has been demonstrated that scoring functions often fail to recognize pose decoys, i.e., ligand poses that are geometrically different from the native binding orientation of a ligand in the experimentally determined crystallographic structure of the protein-ligand complex, but score better than the native pose. In addition, known non-binders may also score better than true binders the former compounds are then designated as binding decoys.[103] Obviously, the presence of both binding and geometrical pose decoys in an ensemble of compound poses resulting from computational docking studies will decrease the accuracy of structure based VS. Moreover, structure-based scoring functions are well-known for having inconsistent VS performances across diverse targets.[47]

Several recent studies have shown that inclusion of pose decoys in the training sets of native structures helps in tuning the scoring functions against decoys, which enhances the accuracy of virtual screening.[63-69] Besides, some other studies have demonstrated that target-specific customized scoring functions[61, 62] are effective methods for improving the discrimination between true ligands and binding decoys in VS for the aimed target. In the present study, we devise a target-specific *pose* (-scoring) filter that is trained to distinguish

native-like poses from pose decoys. The pose filter is developed by applying novel chemical descriptors of the protein/ligand interface and a machine learning classifier to discriminate native-like poses from pose decoys in an ensemble of poses. The training set is generated by multiple rounds of docking of a single cognate ligand to its binding target. Furthermore, we develop a two-step protocol for target-specific virtual screening based on pose filter and MeduaScore. In the first step our pose filter is used to eliminate/penalize putative pose decoys for every ligand, and in the second step the remaining putative native-like poses are scored with physical force field based MedusaScore.[163]

We test the performance of this novel, two-step VS protocol on several benchmark sets available from the Directory of Useful Decoys (DUD).[164] DUD is a specially designed data set including multiple targets, their known ligands, and binding decoys, i.e., compounds that are chemically dissimilar to known ligands but score as well as (or better than) native ligands by the majority of current scoring functions. The recently refined DUD data sets include only lead-like compounds and have the true ligands clustered, making it a ideal benchmark set for testing scaffold hopping capability of VS methods. We use Fred (OpenEye Scientific Software)[36] to dock ligands to target structures and generate poses of each compounds. We find that for most targets eliminating/penalizing pose decoys with the pose filter leads to significant improvement in the enrichment of virtual screening hits, as compared with using the MedusaScore scoring function alone. We compare the VS performance of several popular structure-based scoring functions (XSCORE::HMSCORE[50], Fred::ChemScore[51], Fred::PLP[165], and Fred::Chemgauss3[166]) and several novel VS methods (FieldScreen[167], FLAP::LBX[168], and FLAP::RBLB[168]) that have been recently reported to achieve good performances on the same DUD data sets. We find that our combined scoring

function outperforms other structure-based scoring functions for majority of the targets. Furthermore, the retrieved ligands are less similar to the cognate ligand in comparison with ligand-based approaches (FieldScreen and FLAP::LBX), and are complementary to the ligands retrieved using the structure-based method (FLAP::RBLB).

Our approach employs protocols that are routinely used in cheminformatics research, e.g., binary quantitative structure activity relationship (QSAR) modeling, with the caveat that we use unconventional descriptors of the protein/ligand interface for pose scoring as opposed to using standard chemical descriptors of compounds. An interesting and unique feature of our approach is that the pose classifier is formally trained to recognize geometrical decoys of each ligand; yet it succeeds in correctly recognizing (and eliminating) most of the binding decoys because they are predicted as geometrical decoys. We then employ the MedusaScore physical force field potential for final ranking of poses that remain after filtering.

Methods employing structure-based and ligand-based VS strategies concurrently are only beginning to emerge in the literature (e.g., refs.[168-171]). However, most studies focus on finding consensus hits between the two approaches. In contrast, the method described in this paper, combines for the first time cheminformatics and physical force field based approaches (as reflected in the title of the paper) to structure-based pose scoring into a two-step hierarchical workflow leading to an improved general protocol for virtual screening that can be applied to a large variety of targets.

4.2 Methods

4.2.1 Selection of Targets and Data Sets

The data sets of true ligands and presumed binding decoys for each target in this study are collected from the publicly available Directory of Useful Decoys (DUD).[164] The

DUD data sets were designed to minimize the physical biases inherent in the benchmarking of virtual screening schemes against different biological targets. Each ligand was matched with 36 binding decoy molecules that resemble the native ligand in physical properties, such as molecular weight, LogP, number of hydrogen bonding groups, and number of rotatable bonds but are distinct from the ligand topologically. In total, the DUD database consists of 40 data sets and each ligand has around 36 binding decoys. Further refinement of the DUD data sets is done recently by applying a lead-like filter (MW < 450, AlogP < 4.5) on both ligands and binding decoys[35] as well as the reduced graph cluster filter on ligands[172]. These two filters are intended to mimic the real-life virtual screening campaign and to reduce the analogue bias inflating enrichment in virtual screening. We employ the entire 13 data sets, each of which includes at least 15 ligand clusters, for our method validation. The detailed information of data sets is shown in **Table 4.1**. Six of the 13 targets belong to the kinase family (CDK2, EGFR, p38, PDGFrb, Src, and VEGFr2), where the majority of known ligands occupy ATP binding region. The remaining targets include the class of metalloenzymes (ACE, PDE5), serine protease (FXa), and several other enzymes (AChe, COX-2, HIVRT, and InhA). In order to compare strictly with other VS methods, we use the protein-ligand complexes provided in the original DUD for pose filter training. For VEGFr2 and PDGFrb targets, the complex structures provided in the DUD data sets are generated by docking ligands to apo protein structures.

4.2.2 Docking Methods for Pose Generation

For each target, we prepare the x-ray structure using utilities on Molprobity[173] server to add and optimize hydrogen atoms while correcting potential misinterpretations of amino acid (asparagine, glutamine, or histidine) terminal flips. The crystallographic water

molecules located inside the binding pocket are removed in order to avoid biases when generating poses of molecules but cofactors (e.g., NAD in 1p44 protein model) or metal atoms (e.g., Zinc in the 1o86 protein model) are preserved if they are important for enzyme to function or are involved in interactions with the cognate ligand.

We employ the docking software, Fred (version 2.2.5) from OpenEye Scientific[36] to generate an ensemble of poses for each compound. The ensemble of poses is generated by enumerating rigid rotations and translations of each conformer within the binding site. The conformers of each compound are generated by Omega (version 2.2.1)[36] based on default parameters and the binding site is defined by a 5 Å grid box centered on the cognate ligand. For kinase targets, it is well-known that a hydrogen bond interaction to the protein hinge residues is necessary for both Type I and Type II kinase inhibitors.[174] Thus, this constraint is applied during pose generation to improve docking accuracy.

We apply default parameters provided by Fred during docking except for the number of output poses. For pose filter construction, we retain up to 1000 top-scoring poses generated by docking a single cognate ligand in order to ascertain the conformational diversity of poses. For virtual screening, the top 30 poses (ranked by the Fred's default scoring function, Chemgauss3) of each molecule are preserved for re-scoring by other scoring functions (e.g., MedusaScore).

4.2.3 Ligands *vs.* Binding Decoys and Native-like Poses *vs.* Pose Decoys.

"Binding decoys" are defined as ligands that do not bind to a specific target experimentally (non-binders) but score as high as (or better than) true ligands. Similarly, we use the terms "pose decoys" to describe the poses generated by docking the cognate ligand against the protein target but score better than native-like poses. In our study, native-like

poses are defined as poses generated from docking process with binding mode similar to the native pose. The similarity between poses and the native pose is often measured using Root Mean Square Deviation (RMSD). For the purpose of pose-filter training, we artificially define a RMSD threshold of 4Å to classify poses into native-like poses and pose decoys. The 4Å threshold is consistent with the observation that there is a gap on the distribution plot (MedusaScore *vs.* RMSD) of poses generated by re-docking the cognate ligand for most targets (**Figure 4.1**, Figure S1).

4.2.4 Novel Descriptors of the Protein-Ligand Interface Based on Conceptual DFT

Earlier, we developed the so called ENTess chemical geometrical descriptors[16] of the protein-ligand interface. These descriptors are obtained by using Pauling electronegativity (EN) as an atomic property and Delaunay Tessellation (Tess) to characterize the protein ligand interface as follows. When applied to protein-ligand complexes represented at the atomic resolution level, Delaunay tessellation partitions the protein ligand interface into an aggregate of space-filling, irregular tetrahedra, with both protein and ligand atoms as vertices. Each Delaunay quadruplet is characterized by its unique four-atom composition, which defines the descriptor type (certainly, the same four-body compositions may occur in different, or even the same, protein/ligand interfaces). Furthermore, for each quadruplet we calculate the sum of En values of the composing atom-vertices, which produces the descriptor value. In the previous study,[16] we used the ENTess descriptors to build successful quantitative structure-binding affinity relationship (QSBR) models for 264 x-ray characterized protein-ligand complexes with known binding affinity; the modeling approach followed our standard model development and validation workflow.[117]

In this study, we have developed and employed novel descriptors that are methodologically similar to ENTess descriptors but are theoretically more rigorous.[154] These new descriptors employ pairwise atomic potentials for the protein-ligand complexes (PL) based on maximal charge transfer (MCT)[152] in place of Pauling electronegativities, called here PL/MCT. The PL/MCT is calculated from the following equation (see also **Figure 4.2**):

$$\mathrm{PL/MCT_m} = \sum_{k=1}^{n} \sum_{p}^{1\sim3} \sum_{l}^{1\sim3} (\mathrm{MCT_p} * \mathrm{MCT_l} / \mathrm{d_{pl}})_k \qquad (1)$$

where $\mathrm{PL/MCT}_m$ is the potential of the *m*-th tetrahedron type (i.e. individual descriptor type); *n* is the number of occurrences of this tetrahedron type in a given pose; *p* is the vertex index of a protein atom, *l* is the vertex index of a ligand atom, and d_{pl} is the distance between a pair of protein and ligand atoms found in the same Delaunay tetrahedron. (Note that Delaunay tetrahedra at the protein-ligand interface can be classified based on the relative content of protein and ligand atoms, i.e., three protein and one ligand atoms, two from each, or one protein and three ligand atoms; this explains the tetrahedral type counts in the second and third sum in Equation 1).

The MCT characterizes the maximal electron flow between the donor and acceptor atoms at the protein-ligand interface. It is derived from the conceptual DFT,[152, 155] which provides a theoretical basis for calculating the PL/MCT descriptors. The MCT is calculated as follows, assuming that the total energy of the system is perturbed by the charge transfer up to the second order:

$$\Delta E = \mu \Delta N + 1/2 \eta \Delta N^2 \qquad (2)$$

where ΔE and ΔN represent energy change and charge transfer, respectively. When the total energy is minimized with respect to the charge transfer, $d\Delta E/d\Delta N = 0$, we have

$$\Delta N_{max} = - \mu / \eta \equiv MCT \qquad (3)$$

where μ and η are the chemical potential (negative of electronegativity) and the chemical hardness respectively, defined by $\mu = (\partial E/\partial N)_v$ and $\eta = (\partial^2 E/\partial^2 N)_v$ with ν representing the external potential formed by the framework of atomic nuclei.

4.2.5 Knowledge-based Pose Scoring Filter

As described in the previous session, we classify the poses generated by docking the cognate ligand against the protein target into native-like poses and pose decoys based on the RMSD threshold. The problem of separating native-like poses *vs.* pose decoys for a molecule can be treated as a binary classification problem where poses are characterized by their protein-ligand interfacial descriptors (e.g., PL/MCT descriptors in this study); this is a standard classification problem addressed in many conventional cheminformatics investigations using QSAR modeling. Accordingly, we apply the models (i.e., pose filter) to poses generated in virtual screening, assuming that "bad" poses are similar (based on structural descriptors) to pose decoys in the modeling set and that the filter predicts them as such. "Bad" poses should include both poses of binding decoys and non-native poses of ligands.

To train this knowledge-based pose scoring function for each target, we retain up to 1000 poses generated by re-docking a single cognate ligand against its respective target (**Figure 4.3**). For the VEGFr2 and PDGFrb target, where the native pose is unavailable (an apo structure and a model structure respectively), the pose with lowest MedusaScore is considered as a native pose for RMSD calculation. This is a reasonable assumption since MedusaScore performs well at the benchmarking exercise in native pose prediction.[163] We

classify the poses based on the 4 Å threshold as either native-like (RMSD ≤ 4Å) or pose decoys (RMSD > 4Å) except for the PDE5 pose set where the gap is observed at 3 Å and 3 Å threshold is therefore used. For the poses from re-docking the cognate ligand of *1ckp* (CDK2), we do not observe a characteristic distribution (as, for example, in **Figure 4.1**). Therefore, we regenerate the poses using MedusaDock[175] instead of Fred.

For each pose, we generate PL/MCT descriptors to characterize its interfacial interactions. The degree of similarity of each pose to the native pose is quantified by the Euclidean distance in the PL/MCT descriptor space. Therefore, the pose distribution of each target's modeling set can be characterized by three parameters: the distance to the native pose in the PL/MCT descriptor space (x-axis), the RMSD value (y-axis), and the MedusaScore (colorbar). It is desirable that poses with lower RMSD value correspond to smaller distances to the native pose in the PL/MCT descriptor space (e.g., **Figure 4.1)**.

If this binary data set with native-like poses and pose decoys is quite balanced (their ratio being less than 2-fold), we randomly exclude 20% poses as the test set and construct models based on the remaining 80% poses. In the case of imbalanced distribution, we downsize the major class by retaining only those poses that are similar to poses in the minor class, where the degree of similarity is assessed by Euclidean distance in the PL/MCT descriptor space. For example, the ACE target has 48 native-like poses and 952 pose decoys, after down-sampling, only 49 pose decoys most similar to the native-like poses are retained for model building and validation (**Table 4.2** and **Figure S4.1**).

In the modeling process we employ the Support Vector Machines (SVM) software implemented in the open-source LibSVM[88] package to build binary classification models (i.e., pose filters). We use all models built from SVM with eligible CV accuracy (i.e., pose filter)

for predicting the poses in the test set and poses generated in virtual screening. For each pose, we calculate a FilterScore, which is the fraction of models that predict it as native-like.

It should be emphasized again that only one cognate ligand is used for each target to develop a pose scoring function. However, due to the generic nature of the PL/MCT chemical descriptors this scoring function can be applied to score poses for all diverse ligands used in docking and VS studies.

4.2.6 Physical Force Field-based MedusaScore Scoring Function

MedusaScore[163] is a physical force field-based scoring function that describes the major physical interactions between proteins and ligands, including van der Waals interaction, salt bridge, hydrogen bonding and solvation. MedusaScore is an extension of the Medusa force field,[176] which was developed originally to describe physical interactions within proteins. The original parameters of the Medusa force field were trained on 34 high-resolution protein crystal structures with diverse sequences. Thus, by default MedusaScore is expected to be transferable and applicable to virtual screening of a variety of chemical compounds. Notably there were no protein-ligand data used in the development of MedusaScore, but it still exhibits remarkable accuracy in both docking pose discrimination and binding affinity prediction.[163] During the pose rescoring by MedusaScore, we turn off van der Waals repulsion because this term has been shown to be sensitive to small deviation in ligand poses.[163] It is safe to remove the term in this case because all steric clashes have already been considered during the generation of docking poses.

4.2.7 Data Fusion of MedusaScore and FilterScore

In order to combine the FilterScore and the MedusaScore, which are of different scales, we utilize normalized Z-scores based on their statistical distributions. We firstly apply the pose filter to the poses in VS, and discard poses that are predicted as pose decoy by all eligible models (i.e., FilterScore = 0). Based on the mean (μ) and standard deviation (σ) of each scoring function, the Z-score is calculated from the raw score (X) using **Equation 4**.

$$Z = (X - \mu)/\sigma \quad (4)$$

If the filter is constructed based on the entire sampling space of poses from re-docking the cognate ligand, we apply the same weight for FilterScore and MedusaScore, and the Z-score for each pose is derived as:

$$Z_{combined} = Z_{MedusaScore} - Z_{FilterScore} \quad (5)$$

We add a minus sign for $Z_{FilterScore}$ so that lower Z-score will correspond to better ranked pose, in consistent with MedusaScore convention. If the filter is constructed based on the poses after the down-sampling procedure, we employ a modified scoring strategy based on the concept of applicability domain.[18] We predict the poses within applicability domain using **Equation 5** and predict the poses out of applicability domain by adjusting the weight of FilterScore by DistScore (**Equation 6**).

$$Z_{combined} = Z_{MedusaScore} - 0.5*(Z_{FilterScore} - Z_{DistScore}) \quad (6)$$

The $Z_{DistScore}$ is the Z-score of each pose based on the its distance to the native pose, the mean, and the standard deviation derived from the distribution of PL/MCT-tess Euclidean distance to native pose of all VS poses. Assuming a normal distribution of VS poses comparable with that of poses from re-docking the cognate ligand, we define VS poses that occupy the space of modeling set are within the applicability domain ($Z_{DistScore} < -1$). This threshold is defined by inspecting the covering space of modeling set for five targets (ACE,

CDK2, COX-2, HIVRT, and VEGFr2 in **Figure S4.1**). The final score for each compound in the combined VS scheme is based on the pose with the lowest sum of Z-scores among all the poses retained for that compound.

4.2.8 Evaluation of Virtual Screening Performance

To examine the overall performance of a method for a target data set in virtual screening, we plot the Receiver Operator Characteristic (ROC) curve. And we calculate the Area Under the Curve (AUC) value at each ROC curve to estimate the average performance of a method throughout the ranked list. On the other hand, to quantify the performance of each method at the early stage for a target data set in virtual screening, we employ the ROC enrichment (ROCE) value. Unlike the conventional enrichment factor (EF) metric, ROCE values are independent of the ratio of binding decoys to ligands in a target data set, making them ideal metrics for comparing different methods.[177] The ROCE value is defined as the ratio of true positive rates to the false positive rates, for a given percentage of binding decoys has been observed (i.e., the slope at each point on the ROC plot). We report ROCE values at 0.5%, 1%, 2%, 5% as suggested[177-179] and adopted in previous publications[167, 168, 180]. The meaning of ROCE value at 1% represents the fold enrichment over random performance. In order to emphasize the retrieval of diverse scaffolds, the above metrics (ROCE and AUC) are modified by applying an arithmetic weight to each ligand (awROCE and awAUC)[181], which is inversely proportional to the size of the cluster it belongs to.

We estimate the uncertainty of awROCE/awAUC values using the statistical bootstrapping procedure.[167] For each ranked list, we randomly exclude 20% of data points and recalculate the awROCE values. This is repeated 10000 times and the standard deviation of awROCE values is used to estimate the error of awROCE. Due to the nature of pose filter,

many true negatives (presumed binding decoys) and some false negatives (ligands) are eliminated in several data sets (e.g., ACE, p38, and etc). For these data sets, we calculate the awROCE values based on the reduced list, resulting in a larger estimated error at the low percentages.

4.2.9 Comparison against Structure-based Scoring Functions, FieldScreen, and FLAP

Several popular structure-based scoring functions, which show good docking pose discrimination and binding affinity prediction in publications[60, 163, 182], are selected to compare against our combined approach. It is intriguing to test the performance of these scoring functions in virtual screening since it has been suggested that scoring functions should be tailored for virtual screening.[64, 105] In total, we have tested five scoring functions including MedusaScore, HMSCORE, Chemgauss3, ChemScore[51], and PLP[165]. HMSCORE belongs to the XSCORE[146] scoring utility. Chemgauss3, ChemScore, and PLP are scoring functions implemented in Fred. All of them belong to the class of empirical scoring function except MedusaScore. Moreover, we also compare our approach with some methods that have been published using the same data set including FieldScreen[167] and FLAP (both LBX and RBLB protocols)[168], FieldScreen[167] and FLAP::LBX[168] are two novel ligand-based virtual screening approaches using grid points derived from the cognate ligand as query; FLAP::RBLB approach[168] utilize grid points generated from protein target biased to the cognate ligand. It should be noted that the binding decoys in DUD are designed to be physically similar to, yet topologically distinct from the true ligands. Any ligand-based approaches applied to this data set might generate optimistic results.

4.2.10 2D Chemical Similarity to the Cognate Ligand

We generate the MACCS structural keys for each compound using MOE software (version 2007.09)[183] under standard protocols, and calculate the Tanimoto coefficient (Tc) as the similarity metric between the cognate ligand and compounds in the screening library.

4.3 Results

4.3.1 Native-like vs. Pose Decoys Classifier

The number of poses used in the construction of the pose filter, along with the model statistics for each target, is shown in **Table 4.2**. We also present the distribution of poses of modeling set for each target set in **Figure S4.1**. Depending on the target, the distribution of the native-like poses and pose decoys are either balanced (AChE, EGFR, FXa, InhA, p38, PDE5, PDGFrb, and Src), or biased towards pose decoys (ACE, CDK2, COX-2, HIVRT, and VEGFr2). The details of modeling techniques to address the imbalanced classes have been described in the Methods. The results show that the overall accuracy for both the training set (internal five-fold CV) and the external test set (external five-fold CV) exceeds 90% for all data sets except ACE, HIVRT, and p38 data sets. We predict the VS poses generated from each data set using the models which have CV accuracy greater than 90% except for the HIVRT data set which has no models with CV accuracy above 90%. In the latter case, a threshold of 80% is applied.

It should be emphasized again that for each target-specific filter, we use only one cognate ligand to generate multiple docking poses for further model building. Nevertheless, the filter is applicable to diverse compounds during VS due to the generality of the chemical descriptors we use to characterize the protein-ligand interface. As demonstrated below, these single-ligand based pose filters can significantly improve the accuracy of virtual screening and true hit selection in combination with the MedusaScore force field.

4.3.2 MedusaScore plus Pose Filter Approach Consistently Improve MedusaScore VS Performance

We compare the VS performance of MedusaScore and MedusaScore plus pose filter. We apply the protocols to all the 13 targets in the DUD clustered data set. We measure the VS performance of the two scoring protocols using the awROC curves (**Figure S4.2**). More specifically, we use awAUC values to measure the overall ligand retrieval of the protocols, and use awROCE values at 1% to measure the ligand retrieval at the early stage of the VS (**Figure 4**a).

We find remarkably improved VS performance over the benchmark set by applying the MedusaScore plus pose filter (i.e., the combined scoring function). For all the 13 targets from the DUD set, the awAUC values from using the combined scoring function are consistently higher than from using MedusaScore alone. The improvements are least significant for target EGFR and VEGFr2, where the awAUC value is improved by about 0.02 in both cases. This is probably due to the fact that using MedusaScore alone already results in high awAUC values for these two targets (0.83 and 0.65, respectively). For the other targets, the average degree of awAUC improvement is 0.15, and we find the most improvements are for target AChE and FXa.

When comparing the awROCE values at 1%, we find the combined scoring function is better than using MedusaScore alone for all targets except Src (**Figure 4.4b**). The improvement of awROCE at 1% is most significant for target PDE5 and PDGFrb. For PDE5, we are not able to retrieve any active ligand using MedusaScore alone (awROCE@1% = 0), but the value is improved to approximately 26.5 fold over the random at 1% after combining MedusaScore with pose filter. The pose filter also improves the ligand retrieval for target PDGFrb (awROCE@1% = 43.18), even though the original awROCE value is already high

(23.49) using MedusaScore alone. In addition, for the two targets (EGFR and VEGFr2) where the least improvement of awAUC is observed, the awROCE values at 1% are also improved significantly.

Therefore, by combining MedusaScore with pose filter, we not only improve the overall VS performance (as measured by awAUC), but also improve the early enrichment (as measured by awROCE values at 1%). The improvement seems to be more pronounced at the early stage, which is a desirable feature because practically often only a small fraction of VS hits will be experimentally tested.

4.3.3 MedusaScore plus Pose Filter Approach *vs.* Other Structure-based Scoring Functions

We also compare the VS performance of our combined scoring function with four popular pose scoring functions, including XSCORE::HMSCORE, Fred::ChemScore, Fred::PLP, and Fred::Chemgauss3. We apply those scoring function on the same docking poses and compare their VS performance at the early screening stage (**Figure 4.5**).

We find that our combined pose scoring function outperforms others for most of the targets. At a false positive rate of 0.5%, the combine scoring function has the highest enrichment for seven out of the 13 targets. In addition, the awROCE values for those targets vary from 21.66 to 86.46. In contrast, other scoring functions have the best performance at no more than 3 targets, with awROCE values vary from 12.07 to 43. We find a similar trend at the 1% level. In this case, our combined scoring function has the highest enrichment for six targets, with awROCE values vary from 22.88 to 43.18, while other scoring functions perform best for at most 3 targets, with awROCE values in the range of 9.67 to 26.56. This

comparison demonstrates that our combined scoring functions have better and more consistent VS performance than conventional scoring functions.

The combine scoring function has the worst performance for target Src. We will analyze the reason of Src failure in Discussion. For this target, using MedusaScore alone gives reasonably good enrichment factor of 24.77, close to that from using ChemScore (25.97). With Src as an exception, the combined scoring function tends to have the best performance on targets where using MedusaScore alone also gives fairly good enrichment.

4.3.4 MedusaScore plus Pose Filter Approach *vs.* Other Novel VS Methods

We select a few recently developed VS methods, for which the benchmark results have been reported on the same DUD Cluster data set. One of the methods available for comparison is FieldScreen[167], which is a ligand-based scoring VS method that utilizes molecular fields derived from the cognate ligand as query. Excellent VS performance has also been reported using FLAP[168] molecular field-derived pharmacophores. For FLAP, we compare with two different VS protocols: FLAP::LBX, similar to FieldScreen, which uses ligand-based molecular field, and FLAP::RBLB, which uses both receptor and co-crystallized ligand structure to derive the pharmacophore query. These methods represent the state-of-art VS methods that has been fully tested the entire DUD clustered set.

The awROC curves of scoring methods for each target are shown in **Figure 4.6** and the awROCE values at each stage are tabulated in **Table S4.6-S4.10**. Out of expectation, we find that the VS performance of each scoring method is target-dependent. Our method has the best retrieval for target HIVRT, p38, and PDGFrb. RBLB has clearly the best performance for target PDE5, Src, and VEGFr2. On the other hand, ligand-based VS methods unquestionably outperform other structure-based methods for target COX-2. A close

examination of the COX-2 data set reveals that around 47% of true ligands belong to the same cluster as the cognate ligand used as query. To further investigate the chemical similarity of retrieved ligands to the cognate ligand from different scoring approaches, we compare the average Tanimoto coefficient (Tc) values of true ligands from top 20 ranking lists (**Table 4.3**). Not surprisingly, we find that ligands retrieved by ligand-based VS methods are chemically more similar to the cognate ligands, as measured by the average Tc values. And the average Tc value of the retrieved COX-2 ligands is 0.88 (e.g., FieldScreen), much higher than the average of those for other 12 targets (0.66). The high degree similarity of COX-2 ligands to the query will definitely bias toward the better performance of any ligand-based VS methods such as FieldScreen and FLAP::LBX methods.

We further compare the early enrichment for our combined scoring function and FLAP::RBLB approach because these two methods seem to have the best VS performance at the early stage (in the 0.5% to 5% range). In addition, both methods take advantage of the 3D structures of the receptor and co-crystallized ligands, albeit using different approaches for VS. We want to identify if the different approaches might result in retrieving different ligands. In fact, we find the two methods seem to be complementary to each other. Among the top 20 hits retrieved by the two methods, we find little overlap of the ligand types (**Figure 4.7**). For example, FLAP::RBLB approach is able to retrieve only one cluster for target p38 and PDGFrb, and two clusters for target ACE. In contrast, the MedusaScore with filter approach can retrieve 4, 5, and 7 clusters, for these three targets ACE, p38 and PDGFrb, respectively. Interestingly, the additionally retrieved ligand clusters do not overlap with those obtained using FLAP::RBLB approach. This is also the case for target VEGFr2, where MedusaScore with filter approach retrieved additional five clusters with no overlap with ligands retrieved

by FLAP::RBLB method. For other targets such as AChE, CDK2, EGFR, HIVRT, and InhA, only a small fraction of the newly retrieved clusters overlaps with the ones from FLAP::RBLB approach. Hence, although both methods used receptor and cognate ligand structures for VS, the resulting performance of FLAP::RBLB approach and our approach seem quite complementary on different targets. Combining the two methods shall result in most diverse ligands among the top hits for VS application.

4.4 Discussions

Ligand dependency. The atom types in PL/MCT descriptors are defined based on their exact chemical names. This implementation makes PL/MCT descriptors fairly sensitive to special interactions (e.g., a tri-fluoro functional group). However, using poses with such interactions to construct pose filter makes it too specific. For example, in the Src data set, the cognate ligand we apply to construct pose filter is ANP, which has a long phospho-aminophosphoric chain uncommon to any lead-like ligands. Unsurprisingly, the pose filter predicts almost everything as pose decoys and the combined scoring function deteriorates the VS performance of MedusaScore against the Src data set. Accordingly, we employ another cognate ligand obtained from *1yol* protein-ligand complex to construct pose filter and apply it to virtual screening. The combined scoring function slightly improves the VS performance of MedusaScore against the Src data set (awROCE@1% = 27.6 *vs.* 25.5; awAUC = 0.66 *vs.*0.62).

Similarly, the combined scoring approach only marginally improves the VS performance of MedusaScore against the COX-2 data set, where the pose filter is constructed based on the cognate ligand with a tri-fluoro functional group. For this case, combining MedusaScore with pure DistScore can easier fish out ligands having the distinctive features

of the query compound than using MedusaScore plus pose filter approach (awROCE@1% = 8.7 *vs.* 4.1; awAUC = 0.67 *vs.* 0.39).

Parent scoring function dependency. Theoretically, the target-specific pose scoring filter can be used in combination with any other structure-based scoring function since the definition of pose decoys is based on the RMSD threshold, independent from scoring functions' output. This combined scoring approach can improve the VS performance by a) eliminating binding decoys recognized by pose filter; b) increasing weight for the ligands favored by pose filter. If the ligands favored by pose filter have relatively poorer scores predicted by the parent scoring function and the high-scoring binding decoys are not completely eliminated, then combining the pose filter and the parent scoring function gives limited improvement. For example, in the CDK2 data set, the Cluster #1, #2, #3, #7, and #8 are favored by both the pose filter and Chemgauss3 but are relatively disfavored by MedusaScore, resulting in better performance of combining pose filter with Chemgauss3 (awROCE@1% = 26.0 *vs.* 14.4; awAUC = 0.84 *vs.* 0.71). Another example is the FXa data set, where combining pose filter with Chemgauss3 has better VS performance (awROCE@1% = 15.4 *vs.* 4.8; awAUC = 0.80 *vs.* 0.72). However, docking programs/scoring functions are well-known for having inconsistent VS performances across diverse targets.[47] Therefore, from the practical point of view, it is more important to improve scoring function performance consistently rather than to achieve ideal results for a few targets. The proposed pose filter is designed to this end.

The judgment of threshold to classify native-like poses and pose decoys. We find that the 4 Å-threshold seems optimal considering the distribution of RMSD values of poses and the pose filter performance in virtual screening. Lowering the threshold results in fewer

native-like poses included, which occupy a smaller portion of descriptor space; this ultimately leads to a smaller applicability domain of the pose filter. As a result, using this pose filter in VS leads to poorer performance compared with using the pose filter built based on the 4 Å-threshold. The PDE5 data set is an exception, where a clear gap around 3 Å RMSD can be observed on the pose distribution plot. In virtual screening against the PDE5 data set, the performance of the combined scoring approach with 3 Å-threshold filter is better than with the filter based on the 4 Å-threshold (awROCE@1% = 26.5 *vs* 17.2; awAUC = 0.75 *vs.* 0.72). Moreover, it should be interesting to include the output of a scoring function into the definition of native-like poses and pose decoys (e.g., to train filter only on those native-like poses and pose decoys that are ranked high by the given scoring function), thus, building filters specifically adjusted for each scoring function.

Virtual screening using MedusaScore in combination with DistScore. As shown in **Figure 4.1**, poses with lower RMSD value correspond to having smaller distances to the native pose in the PL/MCT descriptor space. It can be assumed that, for a given molecule, its likelihood to be a true ligand is directly related to how close its poses to the native pose, which can be reflected by the DistScore. We have been applying DistScore in combination with FilterScore to virtual screening for the data sets, where down-sampling during filter construction is necessary. Therefore, for the proper comparison, we also perform virtual screening against all data sets using MedusaScore in combination with DistScore alone (Figure S2). We find that using pose filter to eliminate/penalize pose decoys in virtual screening can consistently improve MedusaScore VS performance and the MedusaScore plus pose filter approach has the best performance for all data sets except for the outliers mentioned above.

4.5 Conclusions

We have developed an integrated knowledge-based pose (-scoring) filter using concepts frequently employed in cheminformatics research such as chemical descriptors of the protein-ligand interface and machine learning techniques for deriving binary pose classification (native-like *vs.* pose decoys) models. We have combined this novel pose filtering procedure with a recently developed physical force field-based scoring function (MedusaScore) to score the docking poses in virtual screening applications. We validated this combined scoring protocol using the refined subsets (13 targets) from the DUD database. The refined DUD sets consist of only lead-like compounds and ligands are clustered based on the reduce graph algorithm, making them suitable for testing scaffold hopping capability of VS methods. The validation results demonstrated that our method can consistently improve the VS performance of MedusaScore provided that the protein-ligand complex is suitable for filter training. Comparing with other conventional structure-based scoring functions, including XSCORE::HMSCORE, Fred::ChemScore, Fred::PLP, and Fred::Chemgauss3, the combined scoring protocol outperforms in six out of 13 data sets at early stage of VS (1% decoys been screened). Moreover, we found that the retrieved ligands by the combined scoring protocol are chemically more diverse than those by other two ligand-based VS methods (FieldScreen and FLAP::LBX) using the same DUD data sets. Interestingly, we observed that our method is complementary to FLAP::RBLB, which is a high-performance VS method that also utilizes both the receptor and the cognate ligand structures.

Our method demonstrated its ability to achieve good enrichments and perform scaffold hopping, suggesting that it could be applied to virtual screening against novel pharmaceutically relevant protein targets to identify promising leads. In particular, this

method is suitable for protein targets with limited ligand binding data available. A single x-ray protein-ligand complex or, as we have demonstrated for PDGFrb target, a homology protein model with a known binder is enough for constructing a successful target-specific pose filter. Additional improvements can be sought for both the pose (-scoring) filter (e.g., using more than one ligand for training, employing alternative atomic properties or potentials for ENTess-like scoring functions, or incorporating other tetrahedral geometric properties), as well as approaches for the integration of knowledge-based and physical force field-based scoringfunctions.

Figures in Chapter 4

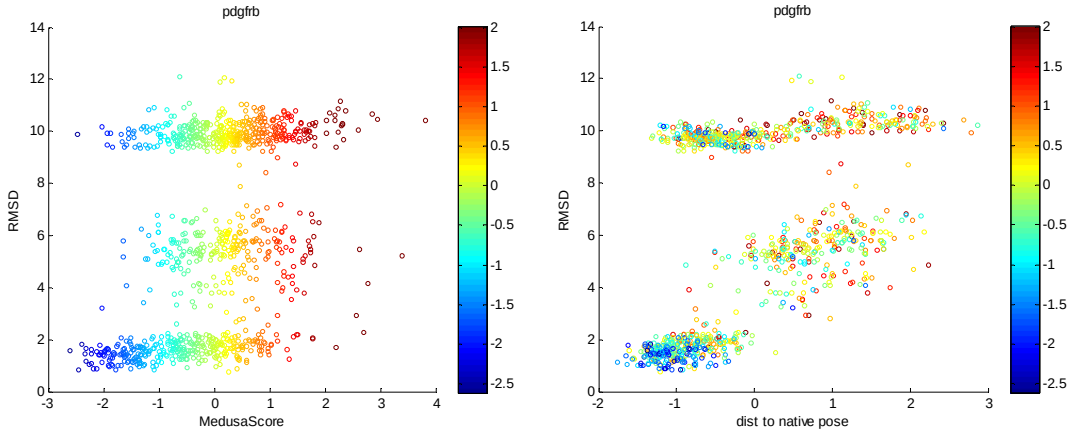


Figure 4.1: The distribution of poses generated by re-docking the ligand structure obtained from the DUD website against the PDGFrb homology protein model.

The pose with the lowest MedusaScore is served as the reference to calculate the RMSD value of poses (the lower MedusaScore values correspond to higher ranks). The left plot shows the pose distribution based on Z-score values of MedusaScore (x-axis) *vs.* RMSD values (y-axis). The right plot shows the pose distribution based on Z-score values of distance to the native pose in PL/MCT descriptor space (x-axis) *vs.* RMSD values (y-axis). The data points are colored corresponding to their Z-score values of MedusaScore**.**

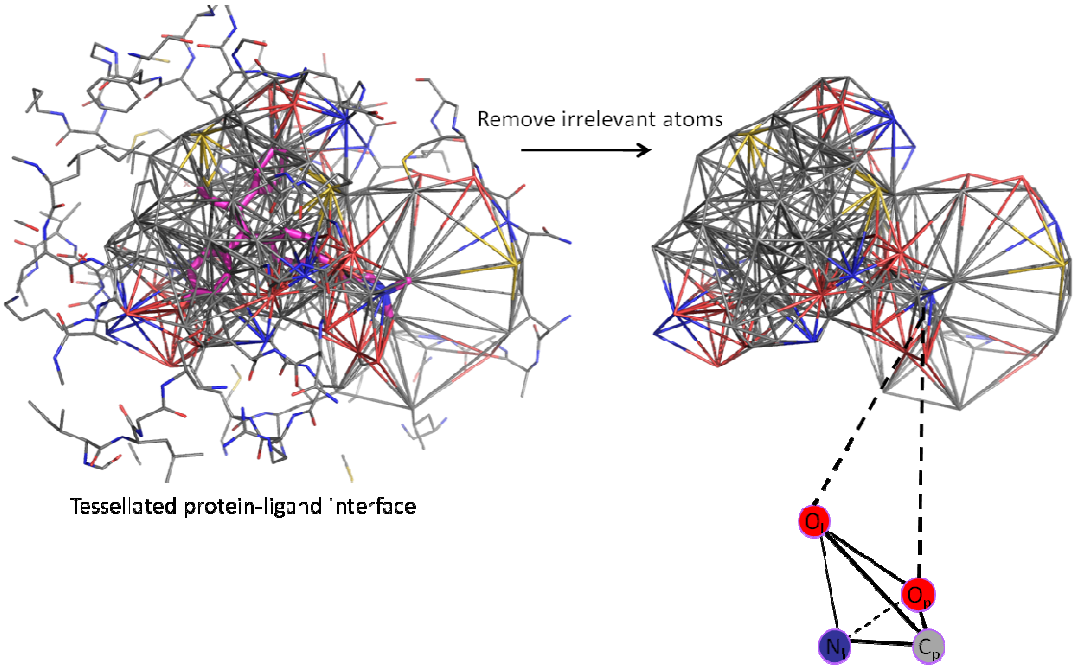


Figure 4.2: Illustration of the method to derive PL/MCT descriptors using the tesselated protein-ligand interface (e.g., 3ERT).

The atom types for protein and ligand are treated differently. For instance, for the tetrahedron at the left corner, C_p and O_p are carbon and oxygen atoms from the protein while O_l and N_l are oxygen and nitrogen atoms from the ligand.

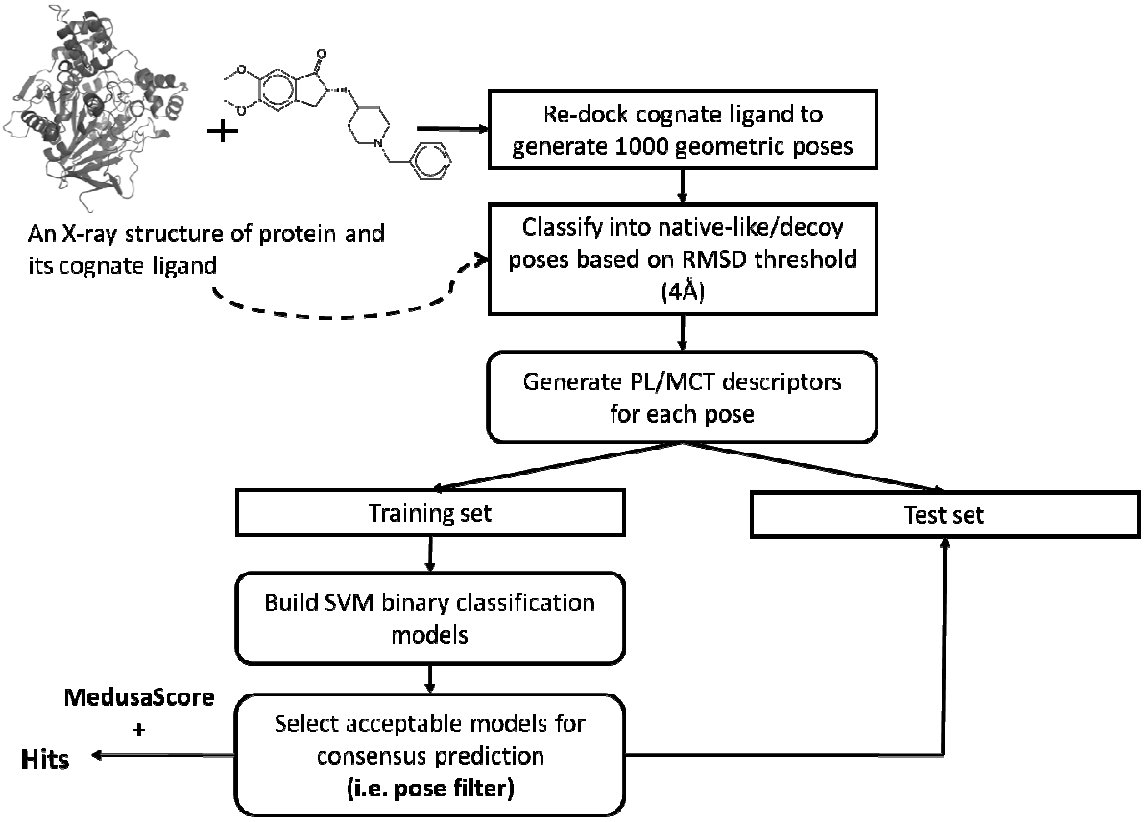


Figure 4.3: Flowchart of the approach described in this paper for developing target-specific pose filters, and their use in combination with MedusaScore for VS.

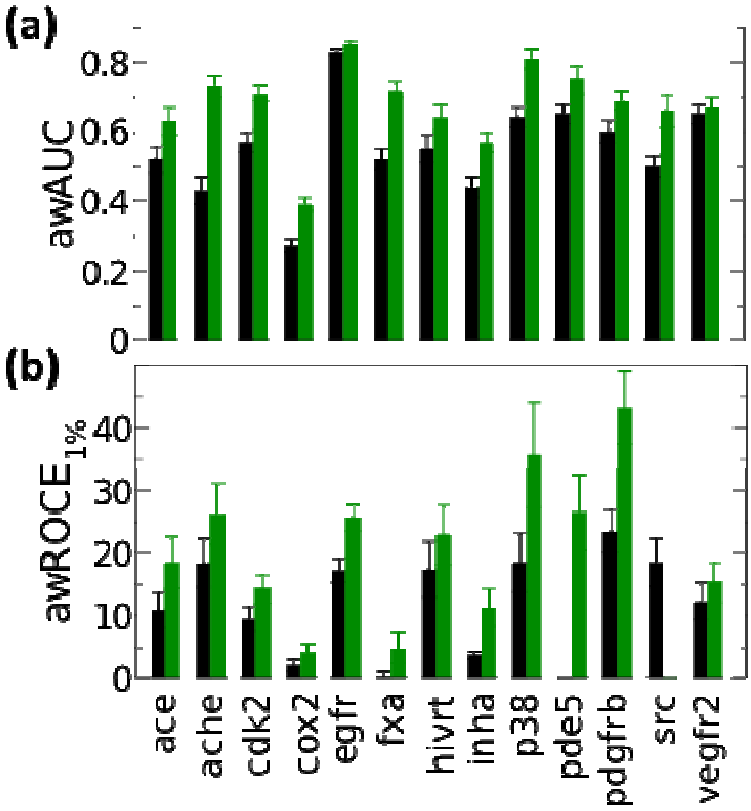


Figure 4.4: The awROCE values at 1% (a) and awAUC values (b) of MedusaScore (black) and MedusaScore + filter approach (dark green) for each target.

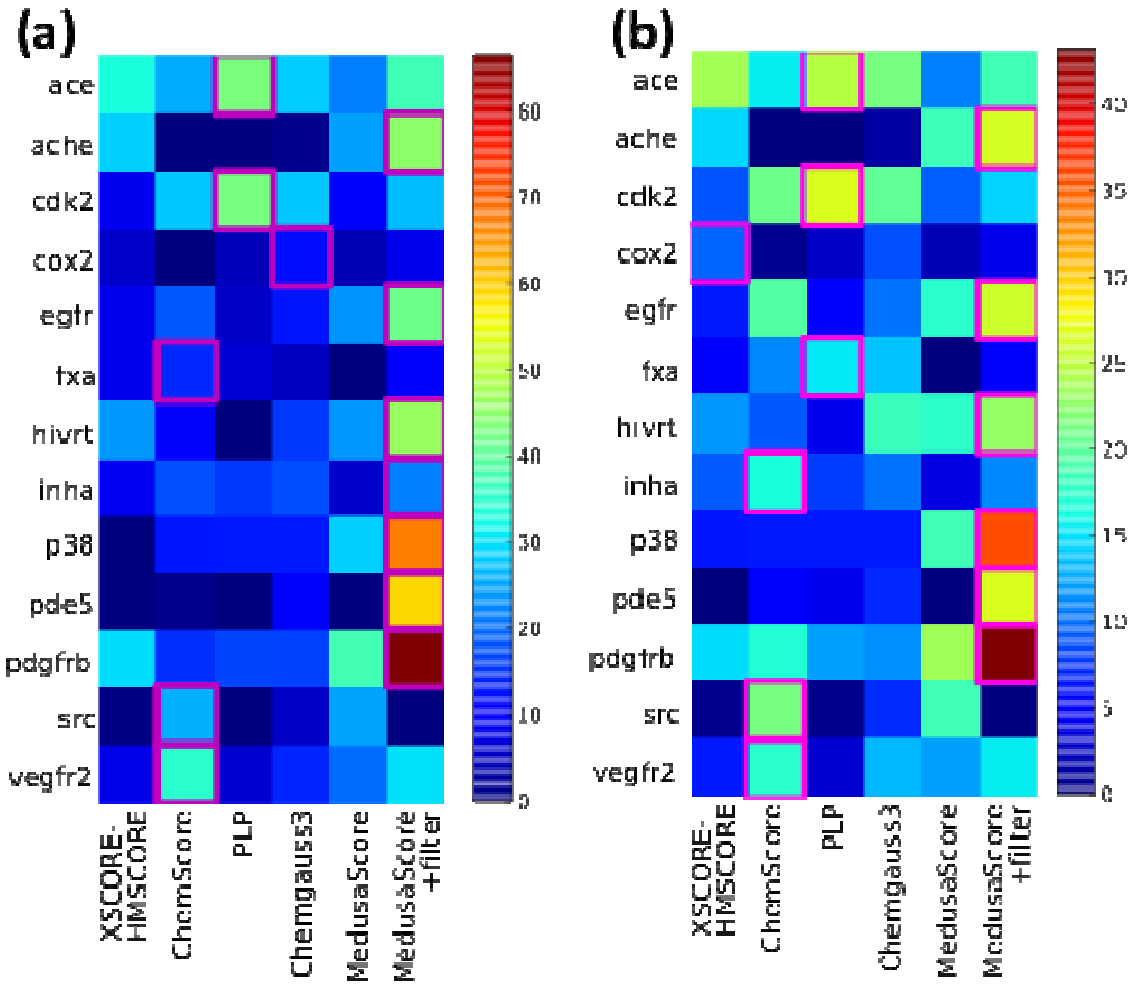
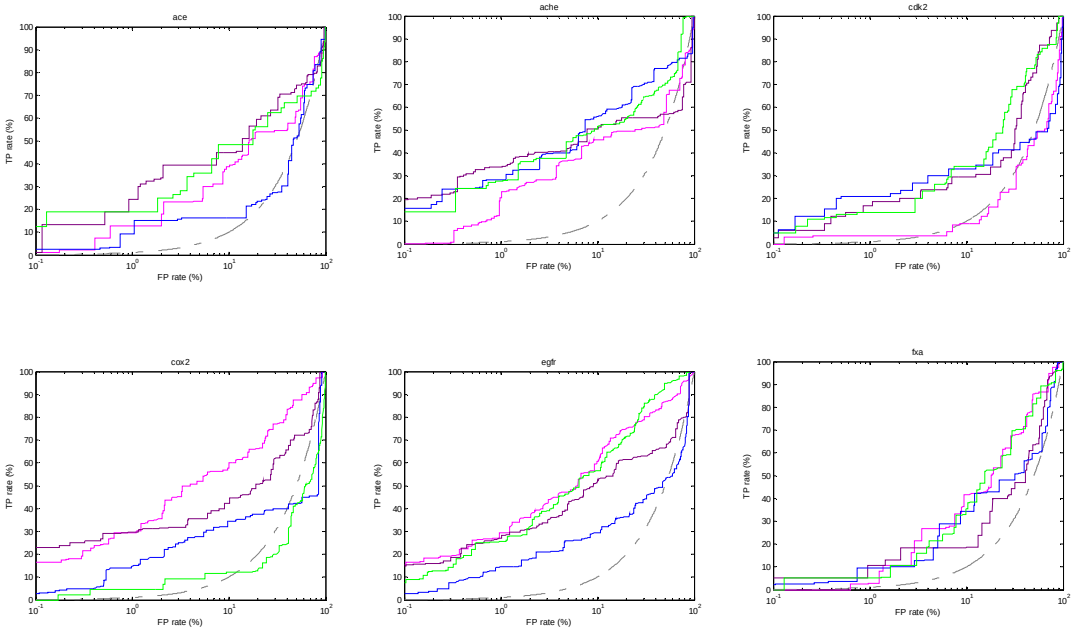


Figure 4.5: The heat map of awROCE values at 0.5% (a) and 1% (b) of several popular structure-based scoring functions (XSCORE::HMSCORE, ChemScore, PLP, Chemgauss3, and MedusaScore) as well as MedusaScore plus Filter approach for each target.

We highlight the highest awROCE values of a scoring method against a particular target (purple box)


ace
ache
cdk2
cox2
egfr
fxa
TP rate (%)
FP rate (%)

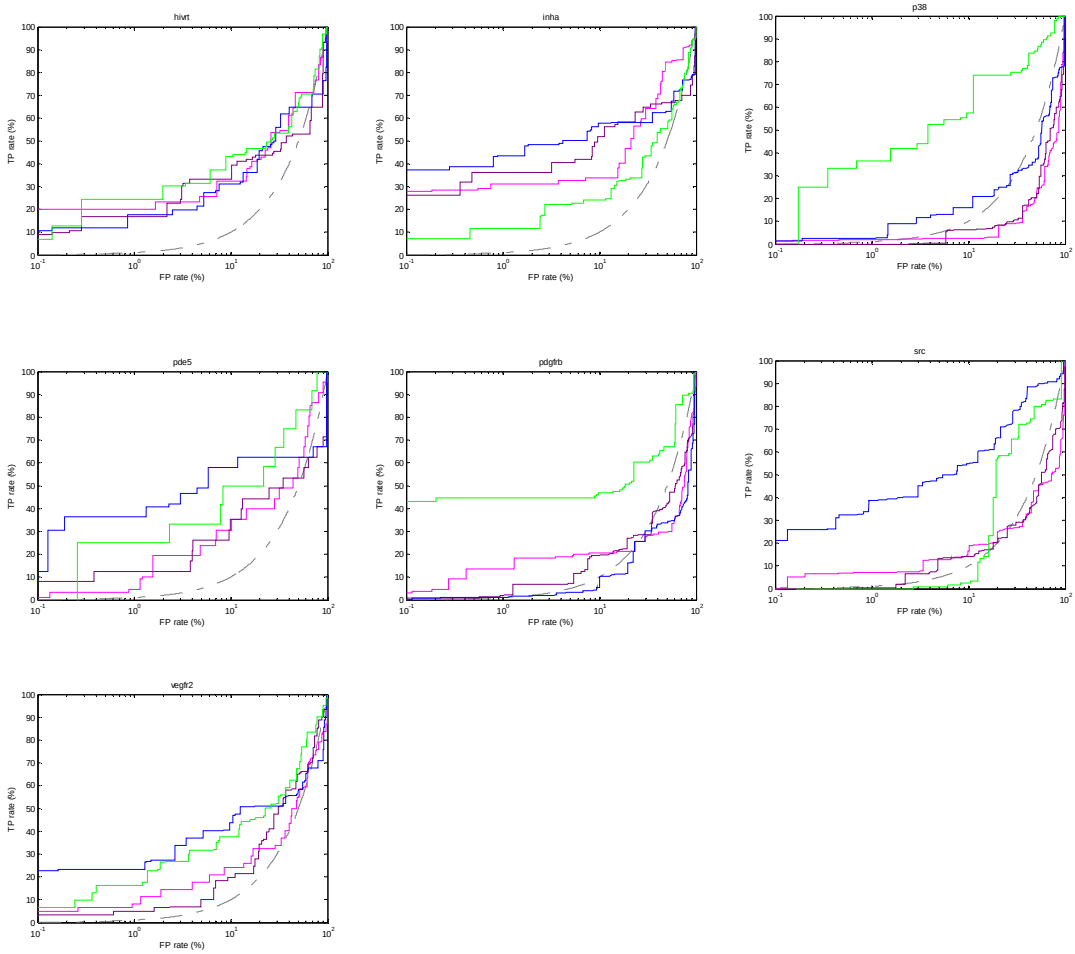


Figure 4.6: The awROC curves of VS experiments for 13 DUD data sets. For each target, the true positive (FP) rate is plotted against the logarithmic false positive (FP) rate.

Gray dot dash lines correspond to the random VS performance, magenta lines are from FieldScreen, purple lines are from FLAP (LBX), blue lines are from FLAP (RBLB), and green lines are from the MedusaScore + pose filter approach

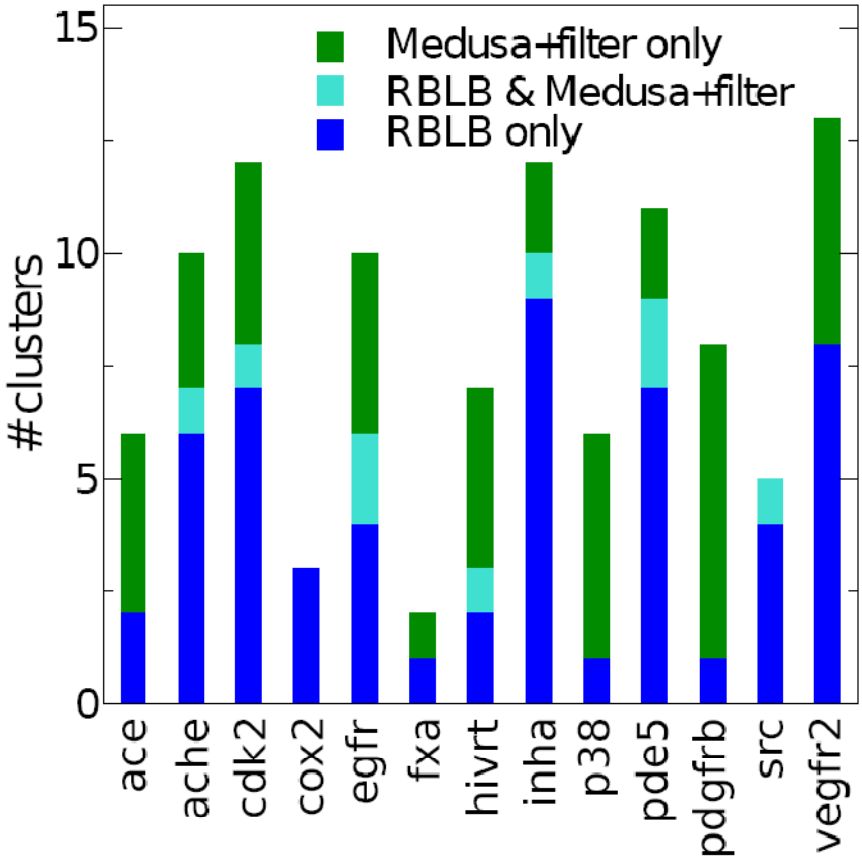


Figure 4.7: The analysis of ligand cluster type retrieval of MedusaScore + filter approach and FLAP::RBLB approach from top 20 ranking list of each data set.

We rearrange the retrieve clusters of each target based on a) the clusters only retrieved by MedusaScore + filter approach (green); b) the clusters only retrieved by FLAP::RBLB approach; c) the overlapping clusters of two approaches (cyan).

Tables for Chapter 4

Table 4.1: Summary of benchmark data sets used in studies described in this paper. The data sets are obtained from DUD website.

Target	Function	PDB	# of ligands	# of decoys	# of clusters
ace	metallopeptidase	1o86	46	1726	19
ache	acetylcholine esterase	1eve	99	3631	19
cdk2	serine/threonine kinase	1ckp	47	1776	32
cox2	cyclooxygenase	1cx2	212	11841	44
egfr	tyrosine kinase	1m17	365	14516	40
fxa	serine protease	1f0r	64	1888	19
hivrt	HIV reverse transcriptase	1rti	34	1415	17
inha	enoyl ACP reductase	1p44	57	2501	23
p38	serine/threonine kinase	1kv2	137	6230	20
pde5	phosphodiesterase	1xp0	26	1562	22
pdgfrb	tyrosine kinase	model[a]	124	5265	22
src	tyrosine kinase	2src	98	5216	21
vegfr2	tyrosine kinase	1vr2[b]	48	2479	31

[a]: protein structure is homology model, the ligand structure is taken from the DUD website

[b]: apo structure, the ligand structure is taken from DUD website

HIV: Human Immunodeficiency Virus; ACP: Acyl Carrier Protein

Table 4.2: Statistics of target-specific pose filters.

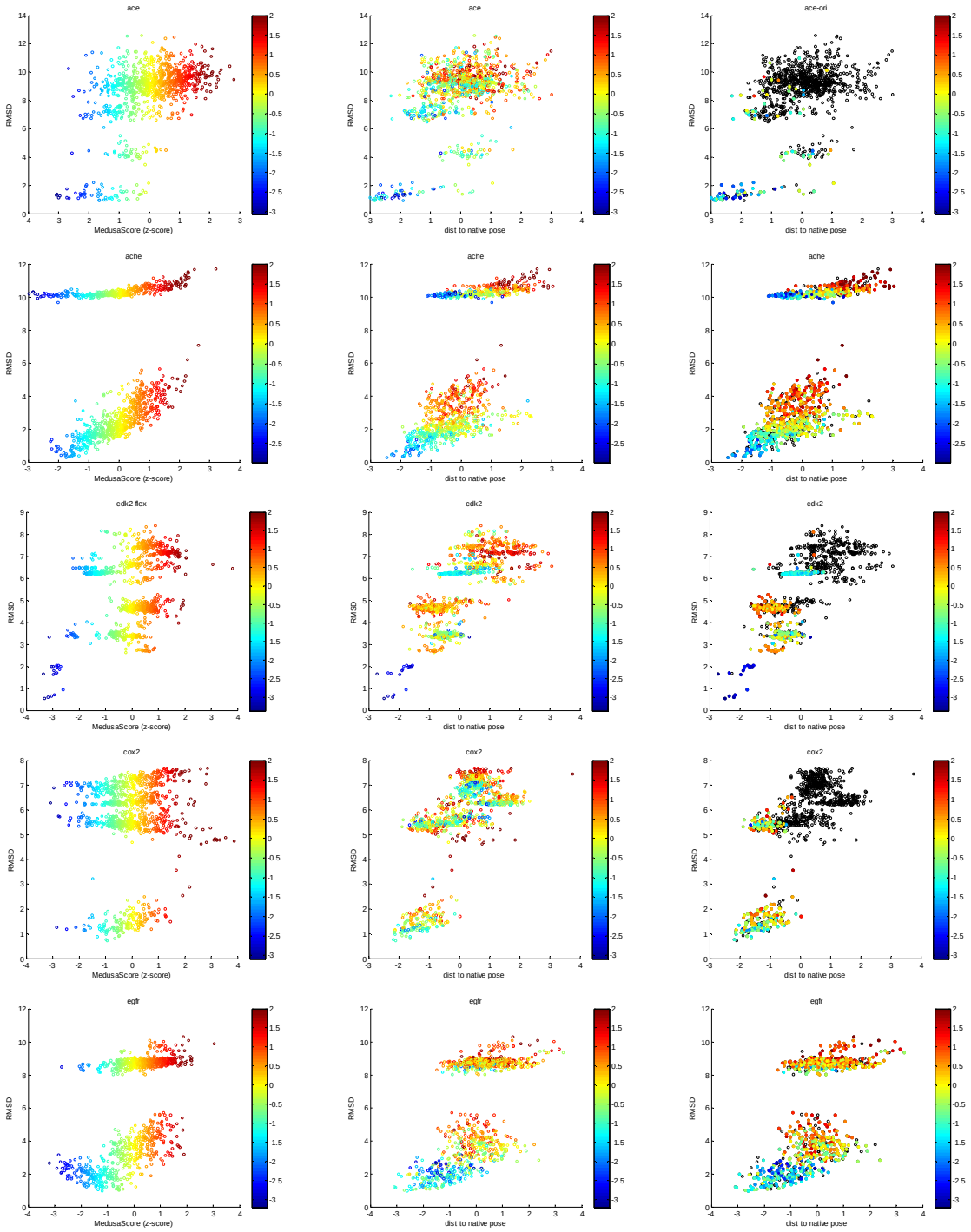
Targets	Training set			Test set		
	Num. native-like poses	Num. pose decoys	CV accuracy[α]	Num. native-like poses	Num. pose decoys	Prediction accuracy
ace	48	49	0.93	13	12	0.89
ache	437	363	0.96	104	94	0.98
cdk2	168	245	0.97	44	60	0.96
cox2	125	96	0.96	36	20	0.99
egfr	296	504	0.94	74	126	0.97
fxa	384	416	0.94	100	100	0.94
hivrt	168	121	0.84	44	29	0.84
inha	296	504	0.96	78	122	0.96
p38	20	24	0.91	6	5	0.82
pde5	295	505	0.96	74	126	0.91
pdgfrb	276	524	0.96	73	127	0.95
src	444	356	0.94	112	88	0.94
vegfr2	132	103	0.93	35	27	0.95

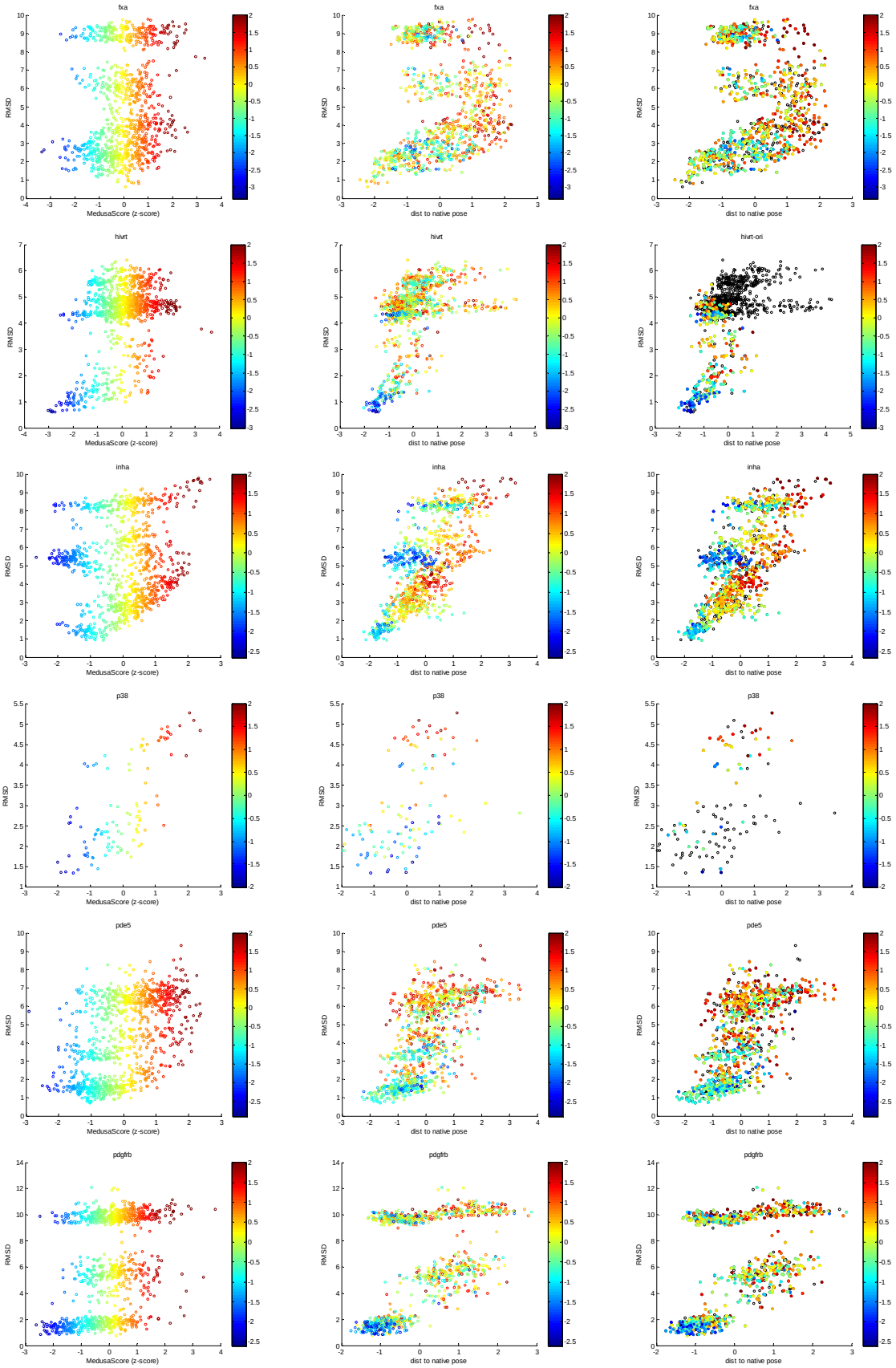
[α]Average CV accuracy is derived from all eligible models with CV accuracy greater than 90% except for the HIVRT data set which has no models with CV accuracy above 90%. Therefore, an 80% threshold is applied

Table 4.3: Average 2D Tc of the active ligands retrieved from the top 20 ranking list of scoring approaches (FieldScreen, FLAP::LBX, FLAP::RBLB, and MedusaScore + filter).

Target	2D similarity			
	FieldScreen	FLAP::LBX	FLAP::RBLB	MedusaScore + Filter
ace	0.75	0.76	0.59	0.74
ache	0.75	0.81	0.52	0.48
cdk2	0.72	0.50	0.49	0.70
cox2	0.88	0.88	0.68	NA
egfr	0.58	0.50	0.45	0.64
fxa	0.49	0.95	0.45	0.45
hivrt	0.78	0.79	0.60	0.59
inha	0.82	0.82	0.69	0.81
p38	0.66	NA	0.45	0.57
pde5	0.74	0.67	0.57	0.69
pdgfrb	0.64	0.64	0.47	0.66
src	0.44	NA	0.47	0.45
vegfr2	0.59	0.67	0.48	0.47
Aver. Similarity	0.68	0.73	0.53	0.60

Supplementary Figures in Chapter 4




fxa
fxa
fxa
hivrt
hivrt
hivrt-ori
inha
inha
inha
p38
p38
p38
pde5
pde5
pde5
pdgfrb
pdgfrb
pdgfrb
RMSD
MedusaScore (z-score)
dist to native pose

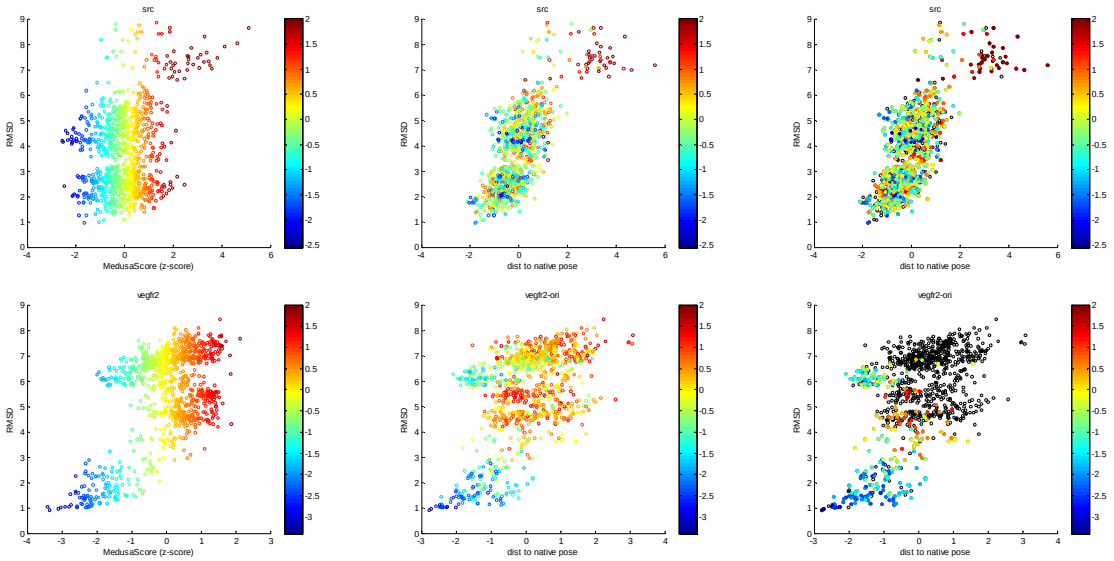


Figure S4.1: The distribution of poses generated from re-docking cognate ligand against its respective target.
The left plot shows the pose distribution based on z-score values of MedusaScore (x-axis) vs. RMSD values (y-axis).
The middle plot shows the pose distribution based on the distance to the native pose (x-axis) vs. RMSD (y-axis).
The right plot highlights the poses used for constructing filter.
The data points are colored corresponding to their z-score values of MedusaScore (the smaller z-score values, the better the MedusaScore).

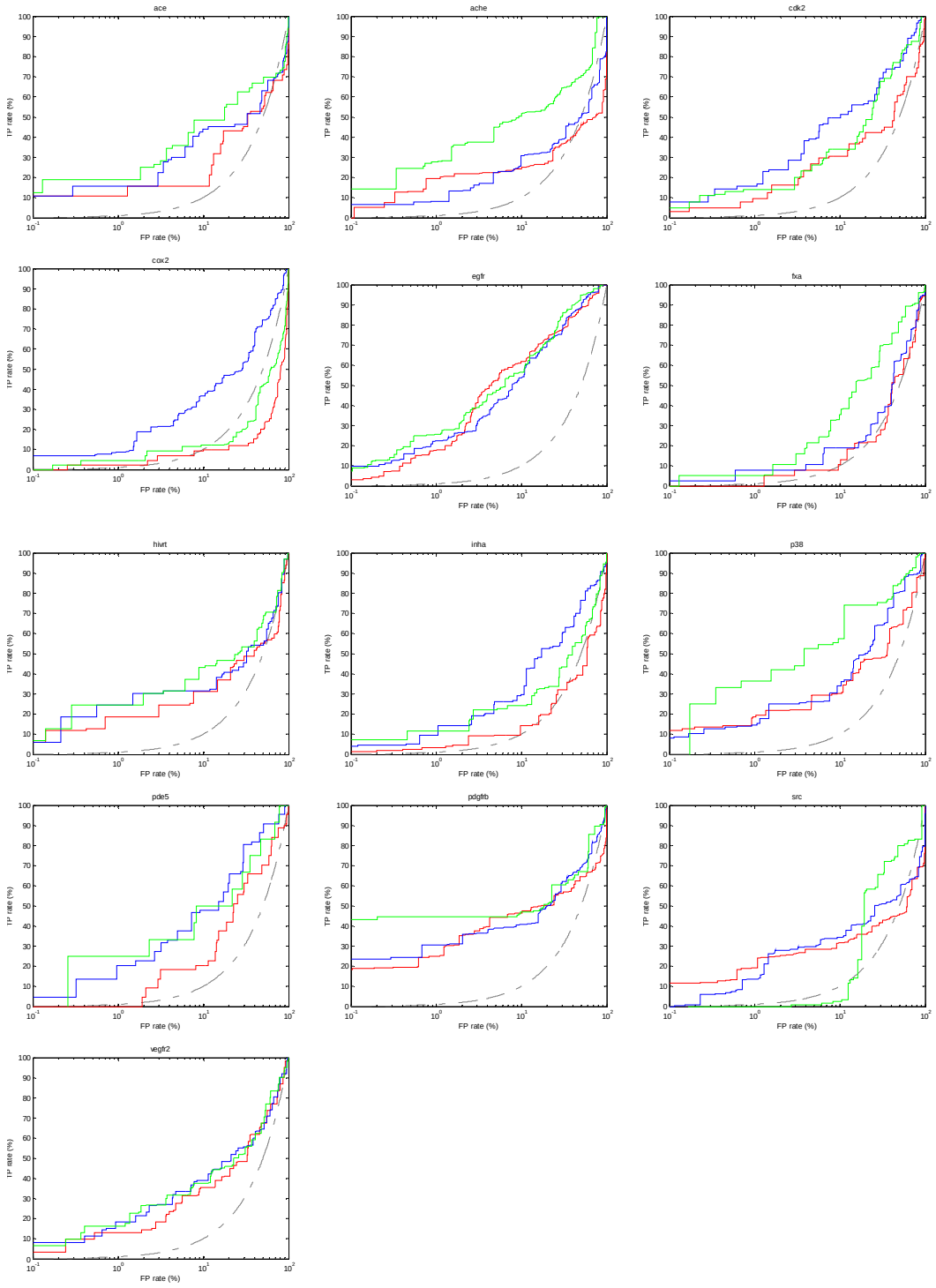


Figure S4.2: The awROC curves of VS experiments for the 13 DUD data sets. For each target, the true positive (FP) rate is plotted against the logarithmic false positive (FP) rate. Gray dot dash lines correspond to the random VS performance, red lines are from

MedusaScore, blue lines are from the MedusaScore +dist.Score approach, and green lines are from the MedusaScore + pose filter approach

Supplementary Tables in Chapter 4

Table S4.1: awROCE enrichment at 0.5% of structure-based scoring functions and the combined scoring approach

target	XSCORE::HMSCORE	Fred::ChemScore	Fred::PLP	Fred::Chemgauss3	MedusaScore	MedusaScore + filter
ace	32.78 ± 8.06	25.45 ± 7.70	**43.00 ± 9.06**	28.11 ± 9.13	21.65 ± 5.53	36.95 ± 8.26
ache	28.48 ± 6.55	0.00 ± 0.00	0.00 ± 0.00	1.12 ± 0.48	24.38 ± 6.74	**44.97 ± 10.16**
cdk2	8.51 ± 4.13	27.71 ± 7.42	**42.88 ± 6.83**	27.90 ± 9.02	11.04 ± 3.19	26.73 ± 4.85
cox2	5.89 ± 1.19	0.00 ± 0.00	4.45 ± 2.10	**12.07 ± 2.49**	4.16 ± 2.07	8.29 ± 2.87
egfr	8.21 ± 1.98	18.38 ± 3.19	5.63 ± 1.70	12.80 ± 2.91	23.59 ± 4.27	**41.61 ± 4.49**
fxa	8.18 ± 5.71	**14.52 ± 4.79**	6.70 ± 4.99	4.86 ± 4.04	0.00 ± 0.00	9.69 ± 4.87
hivrt	23.77 ± 7.62	11.33 ± 5.73	0.01 ± 0.39	15.66 ± 5.91	23.92 ± 8.00	**46.03 ± 9.60**
inha	8.79 ± 1.29	17.79 ± 5.89	15.70 ± 4.28	17.23 ± 4.25	5.92 ± 1.35	**21.66 ± 6.24**
p38	0.00 ± 0.15	12.75 ± 6.45	13.03 ± 6.46	13.12 ± 6.39	28.69 ± 7.32	**66.95 ± 17.33**
pde5	0.00 ± 0.00	1.11 ± 3.36	0.06 ± 0.84	9.59 ± 5.05	0.00 ± 0.00	**59.08 ± 13.50**
pdgfrb	29.65 ± 5.32	15.16 ± 4.12	16.58 ± 4.07	16.26 ± 6.02	37.18 ± 6.51	**86.46 ± 11.56**
src	0.50 ± 0.42	**25.97 ± 7.23**	0.00 ± 0.00	5.62 ± 1.26	24.77 ± 6.29	0.00 ± 0.00
vegfr2	7.85 ± 4.38	**34.53 ± 6.75**	6.18 ± 3.11	14.18 ± 5.57	19.97 ± 5.94	30.00 ± 6.66

Table S4.2: awROCE enrichment at 1% of structure-based scoring functions and the combined scoring approach

target	XSCORE::HMSCORE	Fred::ChemScore	Fred::PLP	Fred::Chemgauss3	MedusaScore	MedusaScore + filter
ace	23.71 ± 4.24	15.47 ± 4.77	**24.56 ± 4.15**	21.33 ± 4.50	10.91 ± 2.86	18.44 ± 4.13
ache	14.64 ± 3.31	0.00 ± 0.00	0.02 ± 0.08	1.37 ± 0.38	18.37 ± 3.91	**26.22 ± 4.95**
cdk2	8.99 ± 2.42	20.90 ± 4.84	**26.56 ± 4.03**	20.13 ± 4.57	9.47 ± 2.06	14.36 ± 2.33
cox2	**9.67 ± 1.85**	0.83 ± 0.51	2.79 ± 1.07	8.79 ± 1.64	2.08 ± 1.03	4.14 ± 1.44
egfr	6.46 ± 1.34	19.48 ± 2.33	4.75 ± 1.34	10.37 ± 1.78	17.22 ± 1.99	**25.69 ± 2.13**
fxa	4.86 ± 2.42	11.26 ± 2.73	15.20 ± 3.76	13.82 ± 4.35	0.15 ± 0.95	4.85 ± 2.44
hivrt	11.94 ± 3.90	9.12 ± 3.49	4.13 ± 3.35	18.13 ± 5.15	17.39 ± 4.45	**22.88 ± 4.77**
inha	9.28 ± 2.15	**16.46 ± 3.17**	7.96 ± 2.15	10.33 ± 2.39	3.73 ± 0.63	11.26 ± 2.91
p38	6.28 ± 3.31	6.52 ± 3.22	6.51 ± 3.23	6.56 ± 3.19	18.46 ± 4.71	**35.65 ± 8.41**
pde5	0.00 ± 0.00	4.80 ± 2.39	4.23 ± 2.43	7.24 ± 2.86	0.00 ± 0.00	**26.49 ± 6.07**
pdgfrb	14.82 ± 2.66	16.93 ± 3.20	12.28 ± 2.73	11.59 ± 2.74	23.49 ± 3.46	**43.18 ± 5.77**
src	0.53 ± 0.21	**21.37 ± 4.10**	0.52 ± 0.31	7.27 ± 1.98	18.46 ± 3.93	0.00 ± 0.00
vegfr2	6.47 ± 2.38	**17.25 ± 3.37**	3.09 ± 1.55	13.22 ± 3.34	12.30 ± 2.89	15.39 ± 3.23

Table S4.3: awROCE enrichment at 2% of structure-based scoring functions and the combined scoring approach

target	XSCORE::HMSCORE	Fred::ChemScore	Fred::PLP	Fred::Chemgauss3	MedusaScore	MedusaScore + filter
ace	15.47 ± 3.01	12.00 ± 2.54	**12.72 ± 2.05**	14.39 ± 2.60	7.84 ± 1.78	11.53 ± 2.56
ache	7.72 ± 1.60	0.77 ± 0.38	0.82 ± 0.60	0.94 ± 0.23	10.47 ± 1.89	**17.16 ± 2.60**
cdk2	7.71 ± 1.57	**15.60 ± 1.91**	14.42 ± 1.88	14.97 ± 1.76	8.30 ± 1.33	7.18 ± 1.16
cox2	7.34 ± 1.21	2.07 ± 0.59	3.10 ± 0.79	**8.23 ± 1.10**	1.04 ± 0.52	2.07 ± 0.72
egfr	4.92 ± 0.79	13.08 ± 1.39	4.63 ± 0.78	8.18 ± 0.98	12.46 ± 1.22	**15.78 ± 1.25**
fxa	2.45 ± 1.22	**10.63 ± 1.99**	10.46 ± 1.99	7.72 ± 1.73	2.40 ± 1.23	4.91 ± 1.66
hivrt	8.59 ± 2.20	6.53 ± 2.17	5.59 ± 2.30	10.76 ± 2.24	8.68 ± 2.22	**12.76 ± 2.81**
inha	4.68 ± 1.05	**8.75 ± 1.59**	5.43 ± 1.24	5.47 ± 1.20	2.63 ± 0.31	5.65 ± 1.48
p38	3.25 ± 1.62	3.27 ± 1.61	3.26 ± 1.61	3.28 ± 1.59	11.34 ± 2.21	**20.63 ± 4.47**
pde5	0.00 ± 0.00	4.90 ± 1.72	2.58 ± 1.49	4.53 ± 1.57	1.91 ± 2.01	**14.27 ± 3.61**
pdgfrb	7.56 ± 1.33	11.29 ± 1.98	6.30 ± 1.36	6.16 ± 1.36	16.32 ± 1.92	**21.58 ± 2.89**
src	0.40 ± 0.23	18.72 ± 2.11	0.74 ± 0.21	5.30 ± 0.68	11.84 ± 1.93	0.00 ± 0.04
vegfr2	4.96 ± 1.29	12.06 ± 2.06	2.49 ± 0.96	7.07 ± 1.55	6.85 ± 1.57	**12.70 ± 1.90**

Table S4.4: awROCE enrichment at 5% of structure-based scoring functions and the combined scoring approach

target	XSCORE::HMSCORE	Fred::ChemScore	Fred::PLP	Fred::Chemgauss3	MedusaScore	MedusaScore + filter
ace	8.48 ± 1.00	5.84 ± 0.89	5.44 ± 0.96	**9.18 ± 1.12**	3.13 ± 0.71	7.04 ± 1.00
ache	3.09 ± 0.64	0.64 ± 0.28	4.08 ± 0.72	1.94 ± 0.32	4.35 ± 0.75	**8.39 ± 1.12**
cdk2	3.64 ± 0.62	6.92 ± 0.70	**7.74 ± 0.82**	7.40 ± 0.77	5.30 ± 0.69	5.14 ± 0.69
cox2	**4.94 ± 0.56**	1.61 ± 0.32	2.40 ± 0.40	5.38 ± 0.56	1.24 ± 0.35	1.66 ± 0.39
egfr	3.60 ± 0.37	9.56 ± 0.50	3.62 ± 0.40	4.36 ± 0.43	**10.41 ± 0.49**	9.25 ± 0.57
fxa	1.97 ± 0.72	5.79 ± 0.93	**6.13 ± 0.91**	3.16 ± 0.69	1.54 ± 0.60	4.18 ± 0.91
hivrt	4.32 ± 1.08	5.14 ± 1.03	3.07 ± 0.80	7.21 ± 1.05	4.56 ± 0.97	**5.88 ± 1.01**
inha	2.69 ± 0.57	**5.54 ± 0.76**	3.10 ± 0.46	3.99 ± 0.69	1.97 ± 0.43	4.18 ± 0.75
p38	1.30 ± 0.65	1.84 ± 0.65	2.64 ± 0.74	1.35 ± 0.64	5.85 ± 1.04	**10.49 ± 1.49**
pde5	0.00 ± 0.00	2.95 ± 0.76	2.00 ± 0.77	3.15 ± 0.69	3.57 ± 0.84	**6.90 ± 1.36**
pdgfrb	3.97 ± 0.62	6.68 ± 0.72	4.27 ± 0.75	4.57 ± 0.64	8.33 ± 0.77	**8.63 ± 1.15**
src	0.72 ± 0.10	**9.67 ± 0.77**	0.49 ± 0.12	3.28 ± 0.76	5.49 ± 0.78	0.19 ± 0.11
vegfr2	1.98 ± 0.52	**6.65 ± 0.80**	2.24 ± 0.58	2.92 ± 0.62	5.44 ± 0.75	6.27 ± 0.79

Table S4.5: awAUC of structure-based scoring functions and the combined scoring approach

target	XSCORE::HMSCORE	Fred::ChemScore	Fred::PLP	Fred::Chemgauss3	MedusaScore	MedusaScore + filter
ace	**0.69 ± 0.04**	0.68 ± 0.03	0.64 ± 0.03	0.66 ± 0.04	0.52 ± 0.04	0.63 ± 0.04
ache	0.42 ± 0.03	0.47 ± 0.03	0.51 ± 0.03	0.47 ± 0.03	0.43 ± 0.04	**0.73 ± 0.03**
cdk2	0.60 ± 0.03	**0.78 ± 0.02**	0.61 ± 0.03	**0.78 ± 0.02**	0.57 ± 0.03	0.71 ± 0.02
cox2	0.68 ± 0.02	0.64 ± 0.01	0.61 ± 0.02	**0.73 ± 0.02**	0.27 ± 0.02	0.39 ± 0.02
egfr	0.57 ± 0.01	0.92 ± 0.00	0.67 ± 0.02	0.66 ± 0.01	0.83 ± 0.01	**0.85 ± 0.01**
fxa	0.57 ± 0.03	0.75 ± 0.02	**0.76 ± 0.02**	0.73 ± 0.02	0.52 ± 0.03	0.72 ± 0.02
hivrt	0.53 ± 0.04	0.68 ± 0.04	0.54 ± 0.04	**0.75 ± 0.03**	0.55 ± 0.04	0.64 ± 0.04
inha	0.29 ± 0.03	0.55 ± 0.03	0.45 ± 0.03	0.51 ± 0.03	0.44 ± 0.03	**0.57 ± 0.03**
p38	0.39 ± 0.03	0.42 ± 0.03	0.40 ± 0.03	0.35 ± 0.03	0.64 ± 0.03	**0.81 ± 0.03**
pde5	0.40 ± 0.03	0.70 ± 0.03	0.60 ± 0.03	0.61 ± 0.03	0.65 ± 0.03	**0.75 ± 0.04**
pdgfrb	0.44 ± 0.03	**0.74 ± 0.01**	0.64 ± 0.02	0.63 ± 0.02	0.60 ± 0.03	0.69 ± 0.03
src	0.44 ± 0.02	**0.83 ± 0.01**	0.45 ± 0.02	0.67 ± 0.02	0.50 ± 0.03	0.66 ± 0.05
vegfr2	0.43 ± 0.03	**0.83 ± 0.01**	0.58 ± 0.03	0.74 ± 0.02	0.65 ± 0.03	0.67 ± 0.03

Table S4.6: awROCE enrichment at 0.5% of FieldScreen, FLAP (LBX), FLAP(RBLB), MedusaScore + filter

target	FieldScreen	FLAP (LBX)	FLAP (RBLB)	MedusaScore + filter
ace	13.76 ± 7.42	28.56 ± 8.92	6.53 ± 1.96	**36.95 ± 8.26**
ache	15.88 ± 5.11	**63.00 ± 8.16**	48.39 ± 8.07	44.97 ± 10.16
cdk2	8.56 ± 3.29	27.66 ± 6.17	**41.61 ± 7.53**	26.73 ± 4.85
cox2	50.93 ± 4.67	**57.33 ± 6.00**	16.74 ± 5.59	8.29 ± 2.87
egfr	**48.57 ± 5.22**	47.96 ± 5.40	19.80 ± 2.89	41.61 ± 4.49
fxa	0.90 ± 2.54	9.57 ± 4.97	8.68 ± 1.81	**9.69 ± 4.87**
hivrt	39.99 ± 8.18	33.47 ± 7.44	27.59 ± 4.46	**46.03 ± 9.60**
inha	59.47 ± 7.05	70.23 ± 8.70	**79.98 ± 8.20**	21.66 ± 6.24
p38	3.99 ± 0.74	0.00 ± 0.00	5.56 ± 2.00	**66.95 ± 17.33**
pde5	8.19 ± 3.90	28.35 ± 5.61	**73.64 ± 10.84**	59.08 ± 13.50
pdgfrb	25.91 ± 5.38	1.61 ± 0.96	2.36 ± 0.57	**86.46 ± 11.56**
src	13.09 ± 4.78	0.00 ± 0.04	**63.41 ± 8.08**	0.00 ± 0.00
vegfr2	13.58 ± 4.51	6.34 ± 3.32	**48.22 ± 6.28**	30.00 ± 6.66

Table S4.7: awROCE enrichment at 1% of FieldScreen, FLAP (LBX), FLAP(RBLB), MedusaScore + filter

target	FieldScreen	FLAP (LBX)	FLAP (RBLB)	MedusaScore + filter
ace	12.23 ± 3.51	**21.84 ± 5.58**	11.15 ± 3.79	18.44 ± 4.13
ache	19.69 ± 5.55	**34.58 ± 3.99**	28.66 ± 4.07	26.22 ± 4.95
cdk2	4.27 ± 1.64	18.41 ± 3.28	**22.11 ± 3.18**	14.36 ± 2.33
cox2	30.33 ± 2.54	**30.70 ± 2.77**	14.98 ± 2.00	4.14 ± 1.44
egfr	**28.04 ± 2.55**	27.45 ± 2.75	14.66 ± 1.52	25.69 ± 2.13
fxa	3.36 ± 2.32	8.27 ± 3.81	**9.67 ± 2.48**	4.85 ± 2.44
hivrt	19.89 ± 4.07	16.64 ± 3.70	18.76 ± 3.59	**22.88 ± 4.77**
inha	11.77 ± 2.31	36.51 ± 3.98	**44.46 ± 4.25**	11.26 ± 2.91
p38	32.58 ± 3.81	0.00 ± 0.00	2.82 ± 1.00	**35.65 ± 8.41**
pde5	2.04 ± 0.36	14.13 ± 2.79	**36.76 ± 5.43**	26.49 ± 6.07
pdgfrb	5.70 ± 2.82	2.19 ± 0.67	1.19 ± 0.29	**43.18 ± 5.77**
src	13.00 ± 2.66	0.33 ± 0.18	**37.53 ± 4.42**	0.00 ± 0.00
vegfr2	6.77 ± 2.39	4.92 ± 1.96	**24.12 ± 3.15**	15.39 ± 3.23

Table S4.8: awROCE enrichment at 2% of FieldScreen, FLAP (LBX), FLAP(RBLB), MedusaScore + filter

target	FieldScreen	FLAB (LBX)	FLAP (RBLB)	MedusaScore + filter
ace	8.08 ± 2.84	**17.00 ± 2.70**	7.52 ± 1.74	11.53 ± 2.56
ache	14.14 ± 1.78	**20.30 ± 2.01**	16.68 ± 1.94	17.16 ± 2.60
cdk2	2.13 ± 0.82	10.23 ± 1.62	**11.94 ± 1.68**	7.18 ± 1.16
cox2	**18.71 ± 1.62**	16.28 ± 1.39	9.64 ± 1.07	2.07 ± 0.72
egfr	**17.75 ± 1.40**	15.49 ± 1.27	8.93 ± 0.90	15.78 ± 1.25
fxa	**5.42 ± 1.35**	7.15 ± 2.29	5.14 ± 1.23	4.91 ± 1.66
hivrt	11.77 ± 2.31	8.71 ± 2.14	9.57 ± 1.70	**12.76 ± 2.81**
inha	16.29 ± 1.91	18.26 ± 1.99	**24.51 ± 2.15**	5.65 ± 1.48
p38	1.14 ± 0.20	0.00 ± 0.00	4.41 ± 1.18	**20.63 ± 4.47**
pde5	10.12 ± 2.07	7.05 ± 1.39	**20.61 ± 2.76**	14.27 ± 3.61
pdgfrb	8.51 ± 1.55	3.20 ± 1.04	0.93 ± 0.18	**21.58 ± 2.89**
src	3.55 ± 1.20	1.07 ± 0.64	**19.91 ± 2.03**	0.00 ± 0.04
vegfr2	7.06 ± 1.56	3.38 ± 1.14	**14.14 ± 1.69**	12.70 ± 1.90

Table S4.9: awROCE enrichment at 5% of FieldScreen, FLAP (LBX), FLAP(RBLB), MedusaScore + filter

target	FieldScreen	FLAB (LBX)	FLAP (RBLB)	MedusaScore + filter
ace	4.64 ± 0.96	**7.59 ± 0.98**	3.28 ± 0.70	7.04 ± 1.00
ache	7.57 ± 0.80	**8.71 ± 0.83**	8.11 ± 0.82	8.39 ± 1.12
cdk2	0.85 ± 0.33	5.02 ± 0.69	**6.18 ± 0.75**	5.14 ± 0.69
cox2	**10.49 ± 0.63**	7.32 ± 0.59	5.72 ± 0.54	1.66 ± 0.39
egfr	**9.34 ± 0.54**	8.38 ± 0.57	5.10 ± 0.40	9.25 ± 0.57
fxa	**5.14 ± 0.87**	3.46 ± 0.84	4.29 ± 1.09	4.18 ± 0.91
hivrt	5.14 ± 0.99	6.63 ± 0.96	4.95 ± 0.82	**5.88 ± 1.01**
inha	6.85 ± 0.74	8.10 ± 0.82	**10.21 ± 0.85**	4.18 ± 0.75
p38	0.53 ± 0.08	0.11 ± 0.02	2.61 ± 0.54	**10.49 ± 1.49**
pde5	4.75 ± 0.96	5.49 ± 0.89	**9.92 ± 1.06**	6.90 ± 1.36
pdgfrb	3.56 ± 0.62	1.35 ± 0.43	0.74 ± 0.14	**8.63 ± 1.15**
src	2.40 ± 0.62	2.38 ± 0.69	**9.67 ± 0.85**	0.19 ± 0.11
vegfr2	3.57 ± 0.63	1.89 ± 0.59	**7.69 ± 0.82**	6.27 ± 0.79

Table S4.10: awAUC of FieldScreen, FLAP (LBX), FLAP(RBLB), MedusaScore + filter

target	FieldScreen	FLAP (LBX)	FLAP (RBLB)	MedusaScore + filter
ace	0.64 ± 0.04	**0.69 ± 0.03**	0.53 ± 0.03	0.63 ± 0.04
ache	0.63 ± 0.03	0.62 ± 0.04	**0.74 ± 0.03**	0.73 ± 0.03
cdk2	0.44 ± 0.02	0.68 ± 0.02	0.50 ± 0.03	**0.71 ± 0.02**
cox2	**0.82 ± 0.02**	0.69 ± 0.02	0.49 ± 0.02	0.39 ± 0.02
egfr	0.82 ± 0.01	0.70 ± 0.02	0.55 ± 0.02	**0.85 ± 0.01**
fxa	**0.73 ± 0.02**	0.61 ± 0.03	0.62 ± 0.03	0.72 ± 0.02
hivrt	**0.64 ± 0.04**	0.56 ± 0.04	0.61 ± 0.04	**0.64 ± 0.04**
inha	**0.72 ± 0.02**	0.66 ± 0.03	0.67 ± 0.03	0.57 ± 0.03
p38	0.28 ± 0.02	0.31 ± 0.02	0.45 ± 0.03	**0.81 ± 0.03**
pde5	0.62 ± 0.03	0.55 ± 0.04	0.64 ± 0.05	**0.75 ± 0.04**
pdgfrb	0.40 ± 0.03	0.44 ± 0.02	0.34 ± 0.03	**0.69 ± 0.03**
src	0.39 ± 0.03	0.44 ± 0.03	**0.80 ± 0.02**	0.66 ± 0.05
vegfr2	0.53 ± 0.03	0.59 ± 0.03	0.61 ± 0.03	**0.67 ± 0.03**

Chapter 5 Conclusions and Future Directions

5.1 Applications of Cheminformatics Approaches to Complement Structure-based Drug Design

In Chapter 2, I have discussed two case studies demonstrating that cheminformatics approaches can complement structure-based drug discovery/drug design and identify promising hits by virtually screening molecular libraries. The first case study is prediction of efflux properties (low *vs.* high) for Gram-negative bacteria, by the binary classification QSAR approach with pharmacophore fingerprint descriptors. Bacterial efflux properties are difficult to model by structure-based methods due to the structural complexity of the efflux pump. However, I have successfully constructed QSAR models which show high prediction accuracy in both internal and external validation. After applying the models to virtual screening, many compounds predicted as low-efflux were experimentally confirmed as such. In the future, I propose to conduct a comprehensive descriptor analysis of predictive models to identify discriminative 3D pharmacophoric features that might contribute to the low-efflux property. The identified 3D pharmacophoric features could provide useful information for the design of low-efflux compounds. The success of this project also suggests that the pharmacophore fingerprint-based SVM QSAR modeling protocol could be applied to predict observations arising from a complex mechanism of action, for example, the permeability of antibiotics.

The second case study is differentiation of AmpC β-lactamase binders *vs.* binding decoys using the binary classification QSAR approach. The binding decoys are false positives mispredicted by the conventional structure-based scoring function (the DOCK score in this case). To differentiate them, I have successfully constructed predictive QSAR models based on rigorous internal and external validations. Applying the models to predict false positives and false negatives from high throughput screening, I showed that the models can discard false positives and can rescue false negatives. Besides, the occurrence frequencies of the 2D chemical descriptors in those models were calculated and those 2D chemical descriptors were then ranked based on their respective frequencies. The top-ranked descriptors could provide the information of possible deficiencies of conventional structure-based scoring functions. For example, the nitrile-group-counts descriptor ("nnitrile") has a high rank (#3) in the frequent descriptor analysis and is only found in the structures of binding decoys, suggesting, it might play an important role in the misprediction by the DOCK score. Furthermore, it is possible to try to include the energetic terms from the structure-based scoring function as descriptors to construct binary classification QSAR models. In that way, the identified frequent descriptors might be found to be directly related to the flawed energetic terms that cause misprediction.

The results of these two example studies suggest that at least in some cases, when a sufficient amount of data is available, QSAR modeling approaches could be used to complement structure-based drug discovery/drug design.

5.2 Development of Single-family based QSBAR Models for Lead Optimization

In **Chapter 3**, I have demonstrated the development of a generic *binding* scoring function, which is a collection of QSBAR models. Compared with the previously reported ENTess scoring function, the new binding scoring function is constructed with a larger number of protein-ligand complexes, representing a more diverse set of protein families, and with novel, protein-ligand interfacial descriptors incorporating conceptual DFT atomic properties. Upon the application of global applicability domain, this new binding scoring function shows acceptable prediction accuracy towards the CSAR data set (n=199, R^2: 0.57), which is much better than the prediction results from the ENTess scoring function (n=135, R^2: 0.30).

It will be beneficial to include more protein-ligand complexes with diverse protein families to construct another new set of predictive QSBAR models based on the same protocol, which could bear larger applicability domain. In **Chapter 3**, I have shown that using the QSBAR models constructed based on the larger data set (PDBbind data set + Set2) significantly improves the external prediction accuracy of Set1 compared with using the models constructed from the smaller Set2 alone (R^2: from 0.40 to 0.50). However, aside from the construction of diverse-family based QSBAR models for generic docking purpose, I would like to propose to construct single-family based QSBAR models in the hope of achieving better prediction accuracy. What is even more important is that the single-family based QSBAR models built with protein-ligand interfacial descriptors could address the selectivity issue of ligands between subfamilies, which is difficult to account for when using 2D chemical descriptors alone (as is done in conventional QSAR modeling).

The collection of protein-ligand complexes, where the protein belongs to the kinase family and the ligand is a Type I ATP competitive inhibitor,[174] can be a good starting point

for constructing single-family based QSBAR models due to its relatively large size. The kinase family consists of many well-studied protein targets. Many marketed drugs act as inhibitors against protein targets in this family.[184] Other than discovering novel chemical scaffolds for a new kinase target involved in a disease, researchers are also highly interested in accurately predicting binding affinity of leads (i.e., lead optimization) and the selectivity profile of compounds. Thus, in the future, I propose to construct the kinase-family based QSBAR models for lead optimization.

5.3 Improvement of Pose (-scoring) Filter for Virtual Screening

In **Chapter 4**, I have described the development of the target-specific *pose* (-scoring) filter with the aim to improve the hit enrichment in structure based virtual screening. The pose filter is developed for each target by building binary classification models that can discriminate native-like poses of ligands *vs*. pose decoys. The training set to develop the filter is generated by multiple rounds of docking a single cognate ligand against its binding target, which generates a large sample of docked poses for this ligand that differ in RMSD from the native pose in the x-ray characterized protein-ligand complex. The pose library is divided into native-like poses (typically, those with RMSD less than 4 Å from the native pose) and decoys (RMSD greater or equal to 4 Å). Each pose is characterized by the chemical descriptors of the protein-ligand interface, which are used as independent variables for developing a binary classifier (i.e., the filter) that discriminates native-like from decoy poses. Furthermore, a two-step scoring protocol for target-specific virtual screening is developed. In the first step, the pose filter is used to remove/penalize putative pose decoys for every compound, and in the second step the remaining putative native-like poses are scored with MedusaScore, which is a conventional force-field-based scoring function.

The validation of this scoring protocol based on the DUD data sets, which are designed for benchmarking, has demonstrated that this method can consistently improve the VS performance of MedusaScore and outperforms many of the conventional structure-based scoring functions. The combined scoring protocol also showed its ability to perform scaffold hopping. In the future, this combined scoring protocol could be applied to virtual screening against several pharmaceutically relevant protein targets to identify promising leads. In particular, this method is suitable for protein targets with limited ligand binding data available. A single x-ray protein-ligand complex or, as I have demonstrated in **Chapter 4**, a homology protein model with a known binder is enough for constructing a target-specific pose filter. Moreover, the pose filter should be theoretically able to add upon any structure-based scoring functions to *consistently* improve their VS performance.

To further improve the pose filter, I propose to develop protein-ligand interfacial descriptors with pharmacophoric node types, e.g., hydrogen-bond donor nodes or hydrophobic nodes. As discussed in **Chapter 4.4**, the atom types in current implementation of PL/MCT-tess descriptors are defined based on their exact chemical names. This implementation makes PL/MCT-tess descriptors fairly sensitive to special interactions. However, using poses with such interactions to construct pose filter makes it too specific (e.g., the Src failure case analyzed in **Chapter 4.4**). Besides, chirality of tetrahedra (see tessellation, **Chapter 2.2.2**) can be also included as a descriptive property, which could significantly increase the amount of information for protein-ligand recognition. The new descriptors should be also useful to construct QSBAR models described in **Chapter 3**.

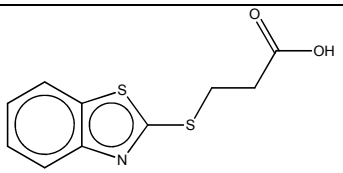
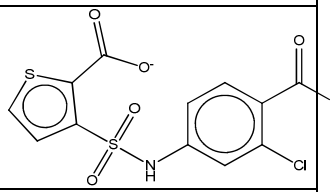
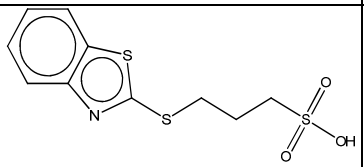
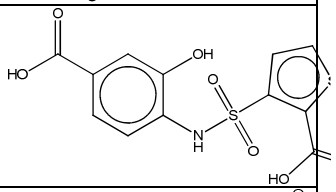
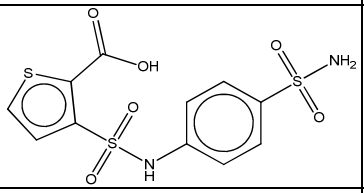
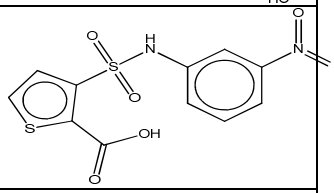
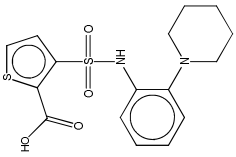
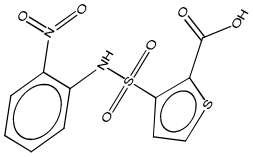
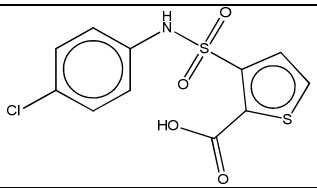
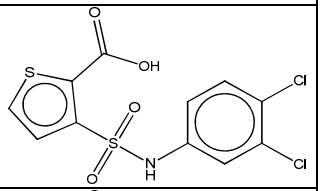
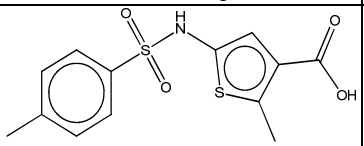
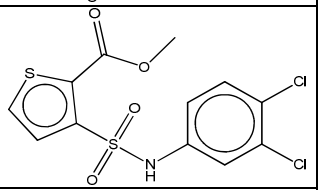
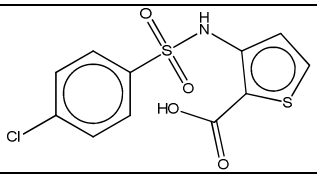
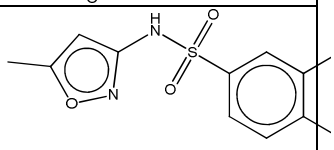
Since the definition of pose decoys is based only on the RMSD threshold, independent from scoring functions' output, theoretically, the pose filter can be used in

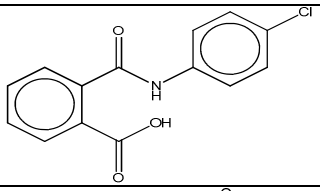
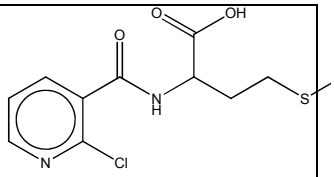
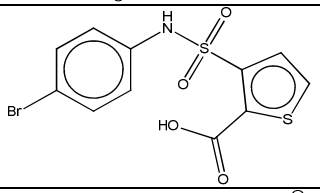
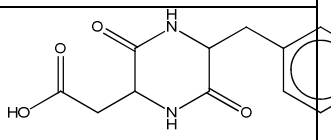
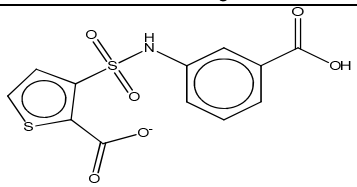
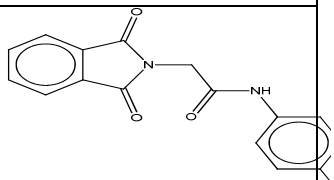
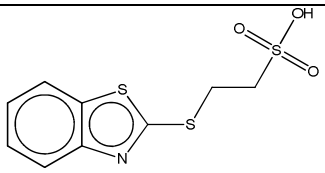
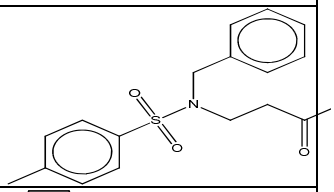
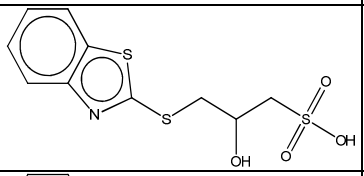
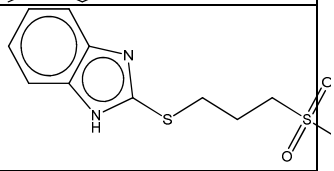
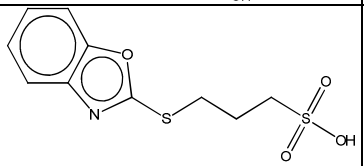
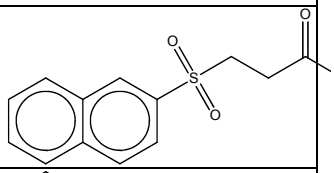
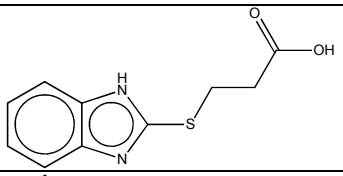
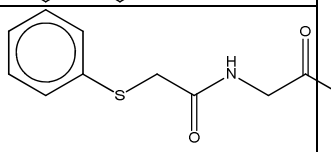
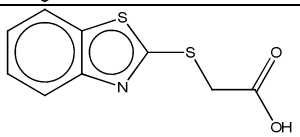
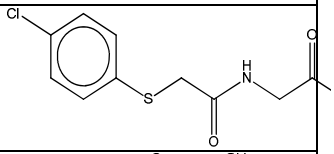
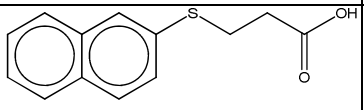
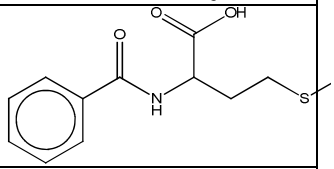
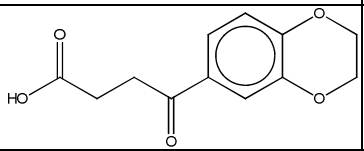
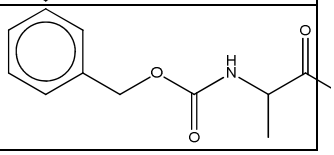
combination with any other structure-based scoring function. However, it should be interesting to include the output of a scoring function into the definition of native-like poses and pose decoys (e.g., to train the pose filter only on those native-like poses and pose decoys that are ranked high by the given scoring function), thus, building filters specifically adjusted for each scoring function. Although this implementation might limit the application of the pose filter to a specific scoring function, it might significantly improve the performance of the combined scoring protocol in virtual screening.

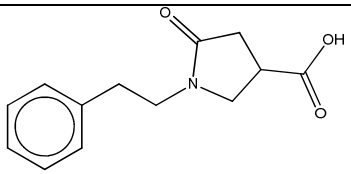
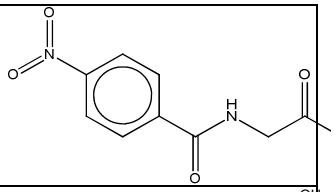
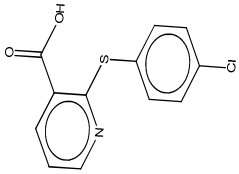
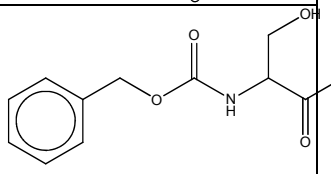
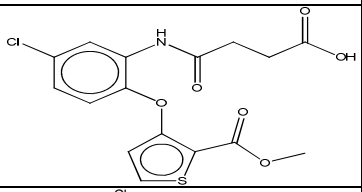
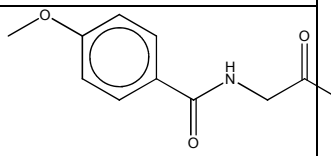
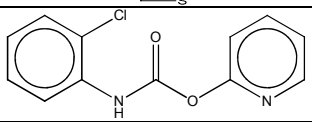
It is a sensible assumption that adding more information should increase the applicability domain of the pose filter. Therefore, the idea of including poses from several protein-ligand complexes to construct a pose filter can be attractive. However, some initial trials suggest that constructing a multi-complex pose filter, by muddling all native-like poses and pose decoys generated from different protein-ligand complexes together, only decreases the discriminative ability of the pose filter when applying it to virtually screen poses generated from docking compounds against a particular protein. It seems that including poses generated from different protein-ligand complexes confuses the pose filter when predicting VS poses from a particular protein. Further investigation of combining poses from several protein-ligand complexes to construct a pose filter is needed. On the other hand, the merge of ranking lists produced by using several different single-complex pose filters is a viable alternative to the multiple-complex pose-filter strategy. Many data fusion techniques (and rank-merge algorithms in particular) can be applied. For example, the Pareto ranking approach generally shows better performance than the most common and simple consensus rank sum approach in the FLAP[168] paper.

Appendices

Appendix I: The AmpC β-lactamase modeling set (16 binders + 25 binding decoys)

Name	Structure	Name	Structure
inhibitor1		inhibitor14	
inhibitor2		inhibitor15	
inhibitor3		inhibitor16	
inhibitor4		inhibitor17	
inhibitor5		inhibitor18	
inhibitor6		inhibitor21	
inhibitor7		nonbinder2	

inhibitor9		nonbinder10	
Inhibitor 11		nonbinder23	
Inhibitor 13		nonbinder30	
Nonbinder 32		nonbinder44	
Nonbinder 33		nonbinder45	
Nonbinder 34		nonbinder46	
Nonbinder 35		nonbinder47	
Nonbinder 36		nonbinder48	
Nonbinder 37		nonbinder51	
Nonbinder 39		nonbinder52	

Nonbinder 40		nonbinder53	
Nonbinder 42		nonbinder54	
Nonbinder 43		nonbinder57	
Nonbinder 77			

Appendix II: The compiled 665 PDBbind data set

PDB code	pKd	PDB code	pKd	PDB code	pKd	PDB code	pKd	PDB code	pKd
10gs.pdb	6.4	1b6k.pdb	8.74	1d09.pdb	7.57	1fkb.pdb	9.7	1ha2.pdb	5.54
1a08.pdb	5.62	1b6m.pdb	8.4	1d3d.pdb	9.09	1fkh.pdb	8.15	1hbv.pdb	6.37
1a0q.pdb	7.57	1b7h.pdb	8.02	1d3p.pdb	7.39	1fki.pdb	7	1heg.pdb	7.74
1a1b.pdb	6.4	1b8n.pdb	10.52	1d4k.pdb	9.22	1fkn.pdb	8.8	1hfs.pdb	8.7
1a1e.pdb	6	1b8o.pdb	10.64	1d4l.pdb	8.77	1fkw.pdb	5.05	1hi4.pdb	4.49
1a30.pdb	4.3	1b9j.pdb	5.96	1d4p.pdb	6.3	1flr.pdb	10	1hih.pdb	8.05
1a42.pdb	9.89	1bcd.pdb	8.7	1d6w.pdb	5.96	1fo0.pdb	5.59	1hk4.pdb	5.31
1a4w.pdb	5.92	1bcu.pdb	3.28	1d7i.pdb	3.6	1fpc.pdb	7	1hpo.pdb	9.22
1a69.pdb	5.3	1bdq.pdb	6.34	1d7j.pdb	3.3	1ftm.pdb	7.61	1hps.pdb	9.22
1a9m.pdb	6.92	1bhx.pdb	6.84	1det.pdb	4.3	1fzj.pdb	8.1	1hvh.pdb	7.96
1abf.pdb	5.42	1bma.pdb	4.59	1df8.pdb	9.7	1fzk.pdb	8.4	1hvi.pdb	10.92
1af6.pdb	1.82	1bn1.pdb	9.34	1dhi.pdb	7.26	1fzm.pdb	7.7	1hvj.pdb	11.4
1afl.pdb	6.28	1bn4.pdb	9.31	1dhj.pdb	6.55	1g2k.pdb	7.96	1hvl.pdb	9.95
1ai4.pdb	2.5	1bnn.pdb	10	1dif.pdb	10.66	1g30.pdb	6.85	1hvr.pdb	9.51
1ai5.pdb	3.72	1bnt.pdb	9.8	1dmp.pdb	9.55	1g35.pdb	8.14	1hwr.pdb	8.33
1ajp.pdb	2.23	1bnu.pdb	9.7	1e1v.pdb	4.92	1g3d.pdb	5.55	1hxb.pdb	9.92
1ajq.pdb	4.31	1bnv.pdb	8.77	1e1x.pdb	5.89	1g45.pdb	8.64	1hxw.pdb	10.82
1ajv.pdb	7.72	1bra.pdb	1.82	1e5a.pdb	7.64	1g46.pdb	8.8	1i9l.pdb	8.48
1ajx.pdb	7.91	1bxo.pdb	10	1e66.pdb	9.89	1g48.pdb	8.41	1i9m.pdb	8.48
1alw.pdb	6.52	1bxq.pdb	7.38	1ejn.pdb	5.62	1g4o.pdb	8.25	1i9n.pdb	8.66
1apw.pdb	8	1c1u.pdb	8.25	1ela.pdb	6.36	1g52.pdb	9.54	1i9o.pdb	8.42
1avn.pdb	3.9	1c1v.pdb	7.64	1elb.pdb	7.15	1g53.pdb	9.04	1i9p.pdb	8.41
1ax0.pdb	3.13	1c5c.pdb	6.96	1eld.pdb	6.7	1g54.pdb	8.82	1i9q.pdb	8.41
1axz.pdb	3.2	1c5p.pdb	4.68	1ele.pdb	6.85	1g7q.pdb	6.06	1if7.pdb	10.52
1b05.pdb	7.12	1c5q.pdb	6.36	1ent.pdb	6.96	1ghw.pdb	4.2	1if8.pdb	9.64
1b11.pdb	7.39	1c5s.pdb	6	1ezq.pdb	9.05	1ghz.pdb	4.8	1iiq.pdb	7.48
1b1h.pdb	7.03	1c5x.pdb	6.68	1f0s.pdb	7.74	1gi6.pdb	6.22	1is0.pdb	7
1b32.pdb	7.1	1c5y.pdb	4.2	1f4e.pdb	2.96	1gi8.pdb	5.05	1iy7.pdb	6.19
1b38.pdb	6.6	1c5z.pdb	4.01	1f4f.pdb	4.62	1gja.pdb	5.42	1j16.pdb	3.84
1b39.pdb	6.92	1c83.pdb	4.85	1f4g.pdb	6.48	1gjb.pdb	6.35	1j17.pdb	5.22
1b3h.pdb	6.21	1c84.pdb	5	1f57.pdb	5.64	1gni.pdb	8.07	1jaq.pdb	4.48
1b46.pdb	5.28	1c86.pdb	4.7	1f5k.pdb	3.74	1gnm.pdb	6.25	1jmg.pdb	6.07
1b4h.pdb	5.46	1c87.pdb	4.2	1fcx.pdb	7.19	1gno.pdb	7.7	1jq9.pdb	8.45
1b51.pdb	7.37	1cbx.pdb	6.35	1fcy.pdb	8.52	1gpk.pdb	5.37	1jqd.pdb	5.16
1b52.pdb	7.12	1ce5.pdb	4.74	1fcz.pdb	9.22	1gz9.pdb	3.51	1jqe.pdb	6.44
1b5h.pdb	6.01	1cim.pdb	8.82	1fd0.pdb	8.4	1h1p.pdb	4.92	1jys.pdb	3.52
1b5i.pdb	7.05	1cin.pdb	8.73	1fh7.pdb	5.24	1h1s.pdb	8.22	1k1i.pdb	6.58
1b5j.pdb	7.43	1cnw.pdb	7.72	1fh8.pdb	6.89	1h22.pdb	9.1	1k21.pdb	8.38
1b6h.pdb	7.82	1cnx.pdb	7.37	1fh9.pdb	6.43	1h23.pdb	8.35	1k22.pdb	8.4

1b6j.pdb	7.92	1cny.pdb	7.85	1fhd.pdb	6.82	1h9z.pdb	5.42	1k4g.pdb	5.85
1k4h.pdb	5.11	1nfy.pdb	8.89	1pb9.pdb	3.62	1sqo.pdb	7.46	1v2j.pdb	3.25
1k9s.pdb	6.52	1nh0.pdb	9.74	1pbq.pdb	6.27	1sqt.pdb	6.2	1v2o.pdb	4.73
1kll.pdb	5.2	1nhu.pdb	5.66	1pph.pdb	5.92	1srg.pdb	5.3	1v2q.pdb	4.13
1kv1.pdb	5.94	1nja.pdb	6.31	1ppk.pdb	7.66	1sri.pdb	6.08	1v2r.pdb	3.55
1kv5.pdb	4.22	1njc.pdb	5.55	1ppm.pdb	5.8	1stc.pdb	8.1	1v2t.pdb	4.71
1kzk.pdb	10.39	1nje.pdb	3.8	1pr5.pdb	3.92	1str.pdb	4.77	1v2w.pdb	4.01
1l2s.pdb	4.59	1nny.pdb	7.66	1pro.pdb	11.3	1sv3.pdb	4.74	1v48.pdb	7.8
1l6m.pdb	8.1	1nvq.pdb	8.25	1pxn.pdb	7.15	1swg.pdb	7.36	1vfn.pdb	5.6
1l83.pdb	3.4	1nwl.pdb	2.39	1pxo.pdb	8.7	1syh.pdb	6.31	1vwl.pdb	5.63
1li3.pdb	4.25	1o0h.pdb	5.92	1pxp.pdb	6.66	1t4v.pdb	7.68	1vwn.pdb	5.82
1li6.pdb	3.8	1o0o.pdb	5.1	1pz5.pdb	5.4	1ta2.pdb	8.52	1vzq.pdb	7.44
1lol.pdb	6.39	1o2h.pdb	7.17	1q63.pdb	5.85	1ta6.pdb	9.13	1w1g.pdb	7.68
1loq.pdb	3.7	1o2j.pdb	6.92	1q7a.pdb	7.19	1tcx.pdb	6.95	1w3j.pdb	6.32
1lor.pdb	11.06	1o2k.pdb	6.92	1q8t.pdb	4.76	1tlp.pdb	7.55	1w4o.pdb	5.22
1lpg.pdb	7.09	1o2o.pdb	6.36	1q8u.pdb	5.96	1tmn.pdb	7.3	1w5v.pdb	8.15
1lpz.pdb	7.6	1o2s.pdb	5.47	1q8w.pdb	5.24	1tnh.pdb	3.37	1w5w.pdb	8.8
1m0n.pdb	2.22	1o2w.pdb	5.85	1qaw.pdb	5.12	1tni.pdb	4	1w5y.pdb	8.48
1m0q.pdb	3.89	1o30.pdb	6.77	1qbu.pdb	10.24	1tnj.pdb	1.96	1ws4.pdb	3
1m2q.pdb	6.1	1o33.pdb	5.74	1qf2.pdb	5.92	1tnk.pdb	1.49	1ws5.pdb	3.03
1m2r.pdb	6.46	1o38.pdb	6.82	1qhc.pdb	7.57	1tnl.pdb	1.88	1wvj.pdb	6.73
1m4h.pdb	9.52	1o3d.pdb	7.13	1qkb.pdb	7.35	1toi.pdb	4.05	1x1z.pdb	11.06
1m7i.pdb	5.4	1o3f.pdb	7.96	1qpb.pdb	1.36	1toj.pdb	3.39	1xd1.pdb	7.92
1mai.pdb	6.68	1o3i.pdb	7.3	1rdi.pdb	2.06	1tok.pdb	2.47	1xgj.pdb	6
1mes.pdb	7.7	1o3j.pdb	6.77	1rdj.pdb	1.66	1trd.pdb	5.4	1xka.pdb	6.88
1meu.pdb	6.1	1o3k.pdb	6.77	1rdl.pdb	2.24	1tsy.pdb	4.96	1xpz.pdb	7.08
1mq6.pdb	11.15	1o3p.pdb	6.66	1re8.pdb	9.52	1ttm.pdb	7.35	1y1m.pdb	1.82
1mrx.pdb	7.26	1obx.pdb	5.72	1rgl.pdb	4.43	1tx7.pdb	4.6	1y1z.pdb	3.08
1msn.pdb	9.09	1ody.pdb	8.1	1rle.pdb	5.8	1tyr.pdb	7	1y3g.pdb	7.4
1mtr.pdb	8.4	1ofz.pdb	4.62	1rnt.pdb	5.19	1u1b.pdb	7.8	1y6q.pdb	11.7
1mu8.pdb	9	1ohr.pdb	8.7	1s38.pdb	5.15	1u2y.pdb	1.74	1yda.pdb	6.55
1mue.pdb	8.64	1oif.pdb	7.72	1s39.pdb	7.7	1u33.pdb	4.6	1ydb.pdb	8.24
1n2v.pdb	4.08	1okl.pdb	6.03	1sb1.pdb	6.89	1ulg.pdb	4.21	1ydd.pdb	7.07
1n5r.pdb	5.66	1ols.pdb	5.82	1sbg.pdb	7.74	1uto.pdb	2.27	1ydr.pdb	5.52
1nc1.pdb	6.12	1olu.pdb	4.41	1sdt.pdb	9.27	1utp.pdb	1.44	1yds.pdb	5.92
1nc3.pdb	5	1om1.pdb	6.77	1sdu.pdb	10.07	1uwt.pdb	5.97	1ydt.pdb	7.32
1ndw.pdb	5.23	1os5.pdb	6.85	1sgx.pdb	5.8	1uz1.pdb	6.89	1z1r.pdb	9.22
1ndy.pdb	6.17	1owe.pdb	6.2	1sl3.pdb	11.85	1v0k.pdb	5.1	1z6e.pdb	9.72
1ndz.pdb	8.11	1owh.pdb	7.4	1sle.pdb	6.17	1v11.pdb	3.98	1z71.pdb	9.18
1nfw.pdb	8.96	1p1o.pdb	5.76	1slg.pdb	3.9	1v16.pdb	3.87	1z9g.pdb	5.64
1nfx.pdb	8.52	1p1q.pdb	4.89	1sqa.pdb	9.21	1v1m.pdb	3.94	1zc9.pdb	3.22

1zdp.pdb	5.74	2cer.pdb	9.22	2j4i.pdb	9	6fiv.pdb	8.08	1g32.pdb	6.11
1zky.pdb	6.25	2ces.pdb	7.25	2j77.pdb	4.89	6rnt.pdb	2.37	1g3b.pdb	5.74
1zoe.pdb	7.4	2cet.pdb	8.02	2j78.pdb	6.42	6std.pdb	8.64	1g3e.pdb	5.38
1zog.pdb	7.15	2cf8.pdb	8.1	2j7b.pdb	6.62	6tim.pdb	3.21	1g4j.pdb	8.7
1zoh.pdb	7	2cgr.pdb	7.28	2j7d.pdb	7.13	7abp.pdb	6.46	1ghv.pdb	4.35
1zp8.pdb	8.77	2ctc.pdb	3.89	2j7e.pdb	7.32	7cpa.pdb	13.96	1ghy.pdb	8.1
1zpa.pdb	8.4	2d0k.pdb	5.02	2j7f.pdb	6.35	7hvp.pdb	9.62	1gi4.pdb	7.19
1zs0.pdb	6.15	2d1o.pdb	7.7	2j7g.pdb	7	7std.pdb	10.72	1gi9.pdb	5.22
1zsf.pdb	9.92	2d3u.pdb	6.92	2j7h.pdb	7.19	8abp.pdb	8	1gj6.pdb	7
1zvx.pdb	9.22	2d3z.pdb	6.64	2qwb.pdb	2.74	8cpa.pdb	9.15	1gj8.pdb	6.96
2aoc.pdb	4.89	2drc.pdb	9.89	2qwc.pdb	3.55	11gs.pdb	5.82	1gjc.pdb	6.35
2aod.pdb	5.66	2er6.pdb	7.22	2qwd.pdb	4.85	1a1c.pdb	6.4	1gjd.pdb	5.22
2aoe.pdb	7.62	2er9.pdb	7.4	2qwe.pdb	7.48	1a94.pdb	7.85	1gnn.pdb	5.68
2aou.pdb	7.73	2erz.pdb	5.66	2rkm.pdb	3.9	1aaq.pdb	8.4	1gvw.pdb	6.96
2aqu.pdb	9.32	2f01.pdb	13	2std.pdb	9.85	1afk.pdb	6.62	1gzc.pdb	3.49
2avm.pdb	5.7	2f80.pdb	8.18	2tmn.pdb	5.89	1ai7.pdb	4.09	1hos.pdb	8.55
2avo.pdb	8.85	2f81.pdb	10.52	2usn.pdb	6.51	1ajn.pdb	2.63	1hpv.pdb	9.22
2avq.pdb	4.39	2f8g.pdb	8.7	3gss.pdb	5.82	1apv.pdb	9	1hpx.pdb	11.26
2avv.pdb	9.26	2fai.pdb	6.24	3pcb.pdb	2.4	1b2h.pdb	4.54	1hvs.pdb	10.3
2ayr.pdb	9.29	2fdp.pdb	7.59	3pcc.pdb	3.62	1b3f.pdb	6.89	1iih.pdb	2.89
2azr.pdb	3.64	2flb.pdb	5.74	3pce.pdb	2	1b3g.pdb	6.7	1izh.pdb	7.7
2b1v.pdb	5.74	2fvd.pdb	8.52	3pch.pdb	5.4	1b4z.pdb	5.23	1izi.pdb	6.59
2b7d.pdb	8.7	2fx6.pdb	3.7	3pcj.pdb	7.22	1b6l.pdb	8.3	1j01.pdb	6.47
2baj.pdb	8.4	2fzc.pdb	2.7	3pck.pdb	6.7	1bn3.pdb	9.89	1jmi.pdb	6.06
2bak.pdb	7.43	2fzz.pdb	10.52	3pcn.pdb	3.66	1bnw.pdb	9.08	1kpm.pdb	5.8
2boh.pdb	8.52	2g5u.pdb	8.49	3std.pdb	11.11	1bp0.pdb	5.4	1lgw.pdb	4
2bok.pdb	6.55	2g8r.pdb	3.99	3tlh.pdb	8.82	1c4u.pdb	10.37	1li2.pdb	4.04
2bpy.pdb	7.4	2g94.pdb	9.52	4er1.pdb	6.62	1c5n.pdb	4.7	1lpk.pdb	7.55
2bq7.pdb	7.05	2gh7.pdb	13	4er2.pdb	9.3	1c5o.pdb	3.49	1m0o.pdb	2.31
2br1.pdb	5.14	2gss.pdb	4.94	4fiv.pdb	6.52	1cps.pdb	6.66	1m2p.pdb	6.11
2brb.pdb	4.86	2h3e.pdb	5.7	4std.pdb	10.33	1d4y.pdb	11.1	1met.pdb	9.4
2brm.pdb	5.89	2h4n.pdb	8.7	4tim.pdb	2.16	1e4h.pdb	8.41	1mq5.pdb	9
2bz6.pdb	7.09	2hdq.pdb	1.4	4tln.pdb	3.72	1elc.pdb	6.66	1mrw.pdb	9.7
2bza.pdb	2.8	2hdr.pdb	1.72	4tmn.pdb	10.17	1epo.pdb	7.96	1msm.pdb	10.48
2bzz.pdb	6.43	2hs1.pdb	8.48	5abp.pdb	6.64	1f0r.pdb	7.66	1mu6.pdb	8.38
2c02.pdb	4.04	2hs2.pdb	8.31	5er1.pdb	6.02	1fao.pdb	7.37	1ndv.pdb	5.92
2c3j.pdb	6.18	2i0a.pdb	11.4	5fiv.pdb	7.66	1fiv.pdb	6.59	1njd.pdb	5.57
2cbu.pdb	5.68	2i0d.pdb	12.1	5std.pdb	10.49	1fkg.pdb	8	1nm6.pdb	10.05
2cbv.pdb	5.48	2izl.pdb	6	5tmn.pdb	8.04	1g1d.pdb	9.44	1nt1.pdb	8.89
2ceq.pdb	7.28	2j34.pdb	7.82	6cpa.pdb	11.52	1g2l.pdb	7.24	1nvr.pdb	8.11
1nvs.pdb	7.82	1y20.pdb	5.32	1tng.pdb	2.93	1qbv.pdb	5.39	2f8i.pdb	7.26

1o0f.pdb	5.3	1y6r.pdb	10.11	1tog.pdb	3.22	1r5y.pdb	6.46	2fgu.pdb	9.18
1o0m.pdb	5.15	1z1h.pdb	8.4	1tom.pdb	8.3	1rdn.pdb	1.84	2g00.pdb	9.74
1o2n.pdb	6.09	1z6s.pdb	4.34	1ugx.pdb	5.91	1rej.pdb	8.3	2hb3.pdb	11.35
1o2x.pdb	5.85	1zgi.pdb	5.34	1usn.pdb	7.74	1rgk.pdb	4.31	2j2u.pdb	7.33
1o2z.pdb	6.11	1zsr.pdb	9.82	1utn.pdb	3.49	1sld.pdb	6.57	2j75.pdb	6.65
1o36.pdb	5.96	220l.pdb	3.4	1uwu.pdb	5.98	1sts.pdb	5	2j79.pdb	5.96
1o3h.pdb	7.3	2aog.pdb	6.28	1ux7.pdb	3	1swr.pdb	6.92	3aid.pdb	6.86
1os0.pdb	6.03	2avs.pdb	7.57	1vjj.pdb	5.77	1syi.pdb	5.44	3pcf.pdb	6.05
1owd.pdb	8.2	2bmz.pdb	3.7	1vot.pdb	6.6	1t7j.pdb	8.7	5tln.pdb	6.37
1oyt.pdb	7.24	2bpv.pdb	7.67	1w1d.pdb	6.52				
1pb8.pdb	5.15	2bqv.pdb	8.05	1w7g.pdb	5.1				
1ppl.pdb	8.55	2bt9.pdb	6.19	1w7x.pdb	8.4				
1pzp.pdb	3.31	2c3l.pdb	5.07	1xgi.pdb	4.85				
1qbs.pdb	9.47	2cji.pdb	8.22	1xq0.pdb	6.34				

References

(1) Dimasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *Journal of Health Economics* **2003,** *22*, 151-185.

(2) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery* **2004,** *3*, 711-715.

(3) Diller, D. J. The synergy between combinatorial chemistry and high-throughput screening. *Current Opinion in Drug Discovery & Development* **2008,** *11*, 346-355.

(4) Dove, A. High-throughput screening goes to school. *Nature Methods* **2007,** *4*, 523-529.

(5) Guido, R. V. C.; Oliva, G.; Andricopulo, A. D. Virtual screening and its integration with modern drug design technologies. *Current Medicinal Chemistry* **2008,** *15*, 37-46.

(6) Chin, D. N.; Chuaqui, C. E.; Singh, J. Integration of virtual screening into the drug discovery process. *Mini-Reviews in Medicinal Chemistry* **2004,** *4*, 1053-1065.

(7) Stahura, F. L.; Bajorath, J. Virtual screening methods that complement HTS. *Combinatorial Chemistry & High Throughput Screening* **2004,** *7*, 259-269.

(8) Manly, C. J.; Chandrasekhar, J.; Ochterski, J. W.; Hammer, J. D.; Warfield, B. B. Strategies and tactics for optimizing the Hit-to-Lead process and beyond - A computational chemistry perspective. *Drug Discovery Today* **2008,** *13*, 99-109.

(9) Varnek, A.; Tropsha, A. *Cheminformatics Approaches to Virtual Screening*; RSC: London, 2008.

(10) Willett, P. A bibliometric analysis of the literature of chemoinformatics. *Aslib Proceedings* **2008,** *60*, 4-17.

(11) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010,** *29*, 476-488.

(12) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007,** *13*, 3494-3504.

(13) Hansch, C.; Fujita, T. Rho-Sigma-Pi Analysis . Method for Correlation of Biological Activity + Chemical Structure. *Journal of the American Chemical Society* **1964,** *86*, 1616-&.

(14) *Chemoinformatics in Drug Discovery*; Wiley-VCH, New York: 2005; pp 223-239.

(15) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling* **2010,** *50*, 1189-1204.

(16) Zhang, S. X.; Golbraikh, A.; Tropsha, A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *Journal of Medicinal Chemistry* **2006,** *49*, 2713-2724.

(17) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000,** *40*, 185-194.

(18) Golbraikh, A.; Tropsha, A. Beware of q(2)! *Journal of Molecular Graphics & Modelling* **2002,** *20*, 269-276.

(19) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *Qsar & Combinatorial Science* **2003,** *22*, 69-77.

(20) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des* **2003,** *17*, 241-253.

(21) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996,** *36*, 563-571.

(22) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development,

and 3D database screening: 1. Methodology and preliminary results. *J. Comput. Aided Mol. Des* **2006,** *20*, 647-671.

(23) Evans, D. A.; Doman, T. N.; Thorner, D. A.; Bodkin, M. J. 3D QSAR methods: Phase and Catalyst compared. *J. Chem. Inf. Model.* **2007,** *47*, 1248-1257.

(24) Hsieh, J. H.; Wang, X. S.; Teotico, D.; Golbraikh, A.; Tropsha, A. Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. *J. Comput. Aided Mol. Des* **2008,** *22*, 593-609.

(25) Tang, H.; Wang, X. S.; Huang, X. P.; Roth, B. L.; Butler, K. V.; Kozikowski, A. P.; Jung, M.; Tropsha, A. Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation. *J. Chem. Inf. Model.* **2009,** *49*, 461-476.

(26) Peterson, Y. K.; Wang, X. S.; Casey, P. J.; Tropsha, A. Discovery of geranylgeranyltransferase-I inhibitors with novel scaffolds by the means of quantitative structure-activity relationship modeling, virtual screening, and experimental validation. *J. Med. Chem.* **2009,** *52*, 4210-4220.

(27) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **2004,** *3*, 935-949.

(28) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982,** *161*, 269-288.

(29) RCSB. PDB. http://www.rcsb.org/ . 2007.

(30) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Wang, S. M. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry* **2004,** *47*, 2977-2980.

(31) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Yang, C. Y.; Wang, S. M. The PDBbind database: Methodologies and updates. *Journal of Medicinal Chemistry* **2005,** *48*, 4111-4119.

(32) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *Journal of Computational Chemistry* **1992,** *13*, 505-524.

(33) Rarey, M.; Wefing, S.; Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. *Journal of Computer-Aided Molecular Design* **1996,** *10*, 41-54.

(34) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **1997,** *267*, 727-748.

(35) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: Applications of AutoDock. *Journal of Molecular Recognition* **1996,** *9*, 1-5.

(36) OpenEye Scientific Software. 2008.

(37) Clark, D. E. What has computer-aided molecular design ever done for drug discovery? *Expert Opinion on Drug Discovery* **2006,** *1*, 103-110.

(38) Wlodawer, A.; Vondrasek, J. Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annual Review of Biophysics and Biomolecular Structure* **1998,** *27*, 249-284.

(39) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007,** *47*, 1504-1519.

(40) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry* **2000,** *43*, 4759-4767.

(41) Perez, C.; Ortiz, A. R. Evaluation of docking functions for protein-ligand docking. *Journal of Medicinal Chemistry* **2001,** *44*, 3768-3785.

(42) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003,** *46*, 2287-2303.

(43) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *Journal of Molecular Modeling* **2003,** *9*, 47-57.

(44) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *Journal of Computer-Aided Molecular Design* **2004,** *18*, 333-344.

(45) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins-Structure Function and Bioinformatics* **2004,** *57*, 225-242.

(46) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins-Structure Function and Bioinformatics* **2004,** *56*, 235-249.

(47) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006,** *49*, 5912-5931.

(48) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins-Structure Function and Genetics* **1999,** *37*, 228-241.

(49) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An All Atom Force-Field for Simulations of Proteins and Nucleic-Acids. *Journal of Computational Chemistry* **1986,** *7*, 230-252.

(50) Wang, R. X.; Lai, L. H.; Wang, S. M. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design* **2002,** *16*, 11-26.

(51) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions .1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design* **1997,** *11*, 425-445.

(52) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology* **2000,** *295*, 337-356.

(53) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999,** *42*, 5100-5109.

(54) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* **2002,** *20*, 281-295.

(55) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001,** *41*, 1422-1426.

(56) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model.* **2005,** *45*, 1134-1146.

(57) Guimaraes, C. R. W.; Cardozo, M. MM-GB/SA rescoring of docking poses in structure-based lead optimization. *Journal of Chemical Information and Modeling* **2008,** *48*, 958-970.

(58) Singh, P.; Mhaka, A. M.; Christensen, S. B.; Gray, J. J.; Denmeade, S. R.; Isaacs, J. T. Applying linear interaction energy method for rational design of noncompetitive allosteric inhibitors of the sarco- and endoplasmic reticulum calcium-ATPase. *Journal of Medicinal Chemistry* **2005,** *48*, 3005-3014.

(59) Shivakumar, D.; Williams, J.; Wu, Y. J.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *Journal of Chemical Theory and Computation* **2010,** *6*, 1509-1519.

(60) Cheng, T. J.; Li, X.; Li, Y.; Liu, Z. H.; Wang, R. X. Comparative Assessment of Scoring Functions on a Diverse Test Set. *Journal of Chemical Information and Modeling* **2009,** *49*, 1079-1093.

(61) Antes, I.; Merkwirth, C.; Lengauer, T. POEM: Parameter optimization using ensemble methods: Application to target specific scoring functions. *Journal of Chemical Information and Modeling* **2005,** *45*, 1291-1302.

(62) Seifert, M. H. J. Robust optimization of scoring functions for a target class. *Journal of Computer-Aided Molecular Design* **2009,** *23*, 633-644.

(63) Huang, S. Y.; Zou, X. Q. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *Journal of Computational Chemistry* **2006,** *27*, 1876-1882.

(64) Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 5. Force-Field-Based Prediction of Binding Affinities of Ligands to Proteins. *Journal of Chemical Information and Modeling* **2009,** *49*, 2564-2571.

(65) Seifert, M. H. J. Optimizing the signal-to-noise ratio of scoring functions for protein-ligand docking. *Journal of Chemical Information and Modeling* **2008,** *48*, 602-612.

(66) Betzi, S.; Suhre, K.; Chetrit, B.; Guerlesquin, F.; Morelli, X. GFscore: A general nonlinear consensus scoring function for high-throughput docking. *Journal of Chemical Information and Modeling* **2006,** *46*, 1704-1712.

(67) Teramoto, R.; Fukunishi, H. Structure-based virtual screening with supervised consensus scoring: Evaluation of pose prediction and enrichment factors. *Journal of Chemical Information and Modeling* **2008,** *48*, 747-754.

(68) Teramoto, R.; Fukunishi, H. Supervised consensus scoring for docking and virtual screening. *Journal of Chemical Information and Modeling* **2007,** *47*, 526-534.

(69) Teramoto, R.; Fukunishi, H. Consensus scoring with feature selection for structure-based virtual screening. *Journal of Chemical Information and Modeling* **2008,** *48*, 288-295.

(70) Poole, K. Outer membranes and efflux: the path to multidrug resistance in Gram-negative bacteria. *Curr. Pharm. Biotechnol.* **2002,** *3*, 77-98.

(71) Nikaido, H. Molecular basis of bacterial outer membrane permeability revisited. *Microbiology and Molecular Biology Reviews* **2003,** *67*, 593-+.

(72) Saier, M. H.; Paulsen, I. T. Phylogeny of multidrug transporters. *Seminars in Cell & Developmental Biology* **2001,** *12*, 205-213.

(73) Yu, E. W.; McDermott, G.; Zgurskaya, H. I.; Nikaido, H.; Koshland, D. E. Structural basis of multiple drug-binding capacity of the AcrB multidrug efflux pump. *Science* **2003,** *300*, 976-980.

(74) Koronakis, V.; Sharff, A.; Koronakis, E.; Luisi, B.; Hughes, C. Crystal structure of the bacterial membrane protein ToOC central to multidrug efflux and protein export. *Nature* **2000,** *405*, 914-919.

(75) Murakami, S.; Nakashima, R.; Yamashita, E.; Matsumoto, T.; Yamaguchi, A. Crystal structures of a multidrug transporter reveal a functionally rotating mechanism. *Nature* **2006,** *443*, 173-179.

(76) Murakami, S.; Nakashima, R.; Yamashita, E.; Yamaguchi, A. Crystal structure of bacterial multidrug efflux transporter AcrB. *Nature* **2002,** *419*, 587-593.

(77) Nargotra, A.; Sharma, S.; Koul, J. L.; Sangwan, P. L.; Khan, I. A.; Kumar, A.; Taneja, S. C.; Koul, S. Quantitative structure activity relationship (QSAR) of piperine analogs for bacterial NorA efflux pump inhibitors. *European Journal of Medicinal Chemistry* **2009,** *44*, 4128-4135.

(78) Nargotra, A.; Koul, S.; Sharma, S.; Khan, I. A.; Kumar, A.; Thota, N.; Koul, J. L.; Taneja, S. C.; Qazi, G. N. Quantitative structure-activity relationship (QSAR) of aryl alkenyl amides/imines for bacterial efflux pump inhibitors. *European Journal of Medicinal Chemistry* **2009,** *44*, 229-238.

(79) Bax, B. D.; Chan, P. F.; Eggleston, D. S.; Fosberry, A.; Gentry, D. R.; Gorrec, F.; Giordano, I.; Hann, M. M.; Hennessy, A.; Hibbs, M.; Huang, J. Z.; Jones, E.; Jones, J.; Brown, K. K.; Lewis, C. J.; May, E. W.; Saunders, M. R.; Singh, O.; Spitzfaden, C. E.; Shen, C.; Shillings, A.; Theobald, A. J.; Wohlkonig, A.; Pearson, N. D.; Gwynn, M. N. Type IIA topoisomerase inhibition by a new class of antibacterial agents. *Nature* **2010,** *466*, 935-U51.

(80) Gwynn, M. A Novel Broad Spectrum Class of Antibacterials for Nosocomial and Biothreat Pathogens. Chemical and Biological Defense Science and Technology Conference. 2009.

(81) Pearson, N. D. Tricyclic nitrogen containing compounds and their use as antibacterials. 6-3-0010.

(82) Pearson, N. D.; Miller W.H. antimicrobial compounds. 2010.

(83) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *Journal of Chemical Information and Computer Sciences* **1999,** *39*, 569-574.

(84) Deanda, F.; Stewart, E. L. Application of the PharmPrint methodology to two protein kinases. *Journal of Chemical Information and Computer Sciences* **2004,** *44*, 1803-1809.

(85) Brady, P. G. Ligand-based Design at GSK via pFPs. 232nd ACS National Meeting, San Francisco, 2006. 2006.

(86) Wu, T.-Y.; Yang, Z. Predictive Statistical Model Building for hERG Liability Based on Pharmacophore Fingerprint Descriptors for an Infectious Disease Project in GSK. 237th ACS National Meeting, Salt Lake City, 2009. 2009.

(87) Yang, Z.; Wu T-Y. Accurate Prediction of logD and hERG Activity by Pharmacophore Fingerprint QSAR. 237th ACS National Meeting, Salt Lake City, 2009. 2009.

(88) Chih-Chung Chang and Chih-Jen Lin. LIBSVM : a library for support vector machines. 2001.

(89) Ceccarelli, M.; Ruggerone, P. Physical Insights into Permeation of and Resistance to Antibiotics in Bacteria. *Current Drug Targets* **2008,** *9*, 779-788.

(90) Sharff, A.; Jhoti, H. High-throughput crystallography to enhance drug discovery. *Curr. Opin. Chem. Biol.* **2003,** *7*, 340-345.

(91) Blundell, T. L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* **2002,** *1*, 45-54.

(92) Dessalew, N.; Bharatam, P. V. Identification of potential glycogen kinase-3 inhibitors by structure based virtual screening. *Biophys. Chem.* **2007,** *128*, 165-175.

(93) Lu, I. L.; Huang, C. F.; Peng, Y. H.; Lin, Y. T.; Hsieh, H. P.; Chen, C. T.; Lien, T. W.; Lee, H. J.; Mahindroo, N.; Prakash, E.; Yueh, A.; Chen, H. Y.; Goparaju, C. M.; Chen, X.; Liao, C. C.; Chao, Y. S.; Hsu, J. T.; Wu, S. Y. Structure-based drug design of a novel family of PPARgamma partial agonists: virtual screening, X-ray crystallography, and in vitro/in vivo biological activities. *J. Med. Chem.* **2006,** *49*, 2703-2712.

(94) Zhou, Y.; Peng, H.; Ji, Q.; Qi, J.; Zhu, Z.; Yang, C. Discovery of small molecule inhibitors of integrin alphavbeta3 through structure-based virtual screening. *Bioorg. Med. Chem. Lett.* **2006,** *16*, 5878-5882.

(95) Du, L.; Li, M.; You, Q.; Xia, L. A novel structure-based virtual screening model for the hERG channel blockers. *Biochem. Biophys. Res. Commun.* **2007,** *355*, 889-894.

(96) Kellenberger, E.; Springael, J. Y.; Parmentier, M.; Hachet-Haas, M.; Galzi, J. L.; Rognan, D. Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening. *J. Med. Chem.* **2007,** *50*, 1294-1303.

(97) Zhao, L.; Brinton, R. D. Structure-based virtual screening for plant-based ERbeta-selective ligands as potential preventative therapy against age-related neurodegenerative diseases. *J. Med. Chem.* **2005,** *48*, 3463-3466.

(98) Evers, A.; Klabunde, T. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J. Med. Chem.* **2005,** *48*, 1088-1097.

(99) Oh, M.; Im, I.; Lee, Y. J.; Kim, Y. H.; Yoon, J. H.; Park, H. G.; Higashiyama, S.; Kim, Y. C.; Park, W. J. Structure-based virtual screening and biological evaluation of potent and selective ADAM12 inhibitors. *Bioorg. Med. Chem. Lett.* **2004,** *14*, 6071-6074.

(100) Christmann-Franck, S.; Bertrand, H. O.; Goupil-Lamy, A.; der Garabedian, P. A.; Mauffret, O.; Hoffmann, R.; Fermandjian, S. Structure-based virtual screening: an application to human topoisomerase II alpha. *J. Med. Chem.* **2004,** *47*, 6840-6853.

(101) Kim, Y. G.; Thai, K. M.; Song, J.; Kim, K. K.; Park, H. J. Identification of novel ligands for the Z-DNA binding protein by structure-based virtual screening. *Chem. Pharm. Bull. (Tokyo)* **2007,** *55*, 340-342.

(102) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006,** *46*, 401-415.

(103) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for docking. *J. Med. Chem.* **2005,** *48*, 3714-3728.

(104) Park, H.; Lee, J.; Lee, S. Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins* **2006,** *65*, 549-554.

(105) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, B. S.; Johnson, A. P. eHiTS: an innovative approach to the docking and scoring function problems. *Curr. Protein Pept. Sci.* **2006,** *7*, 421-435.

(106) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* **2002,** *20*, 281-295.

(107) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999,** *42*, 5100-5109.

(108) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001,** *41*, 1422-1426.

(109) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model.* **2005,** *45*, 1134-1146.

(110) Powers, R. A.; Morandi, F.; Shoichet, B. K. Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure.* **2002,** *10*, 1013-1023.

(111) Tropsha, A. Application of Predictive QSAR Models to Database Mining. In *Cheminformatics in Drug Discovery.*; Oprea, T. Ed.; Wiley-VCH: 2005; pp 437-455.

(112) Medina-Franco, J. L.; Golbraikh, A.; Oloff, S.; Castillo, R.; Tropsha, A. Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. *J. Comput. Aided Mol. Des* **2005,** *19*, 229-242.

(113) de Cerqueira, L. P.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model.* **2006,** *46*, 1245-1254.

(114) Oloff, S.; Mailman, R. B.; Tropsha, A. Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J. Med. Chem.* **2005,** *48*, 7322-7332.

(115) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004,** *47*, 2356-2364.

(116) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y. D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of ambergris fragrance compounds. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 582-595.

(117) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007,** *13*, 3494-3504.

(118) NCBI. PubChem. http://pubchem.ncbi.nlm.nih.gov/ . 2007.

(119) Shoichet, B. K. Dr. Brian Shoichet Take-away Webpage. http://shoichetlab.compbio.ucsf.edu/take-away.php . 2007.

(120) Powers, R. A.; Morandi, F.; Shoichet, B. K. Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure.* **2002,** *10*, 1013-1023.

(121) Tondi, D.; Morandi, F.; Bonnet, R.; Costi, M. P.; Shoichet, B. K. Structure-based optimization of a non-beta-lactam lead results in inhibitors that do not up-regulate beta-lactamase expression in cell culture. *J. Am. Chem. Soc.* **2005,** *127*, 4632-4639.

(122) Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* **2007,** *50*, 2385-2390.

(123) PubChem. PubChem Bioassay AID 584. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=584 . 2007.

(124) PubChem. PubChem Bioassay AID 585. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=585 . 2007.

(125) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **2005,** *1*, 146-148.

(126) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Divers.* **2002,** *5*, 231-243.

(127) Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988,** *28*, 31-36.

(128) Tripos. Sybyl. 2007.

(129) eduSoft LC. MolconnZ. 2007.

(130) Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic Press: New York, 1976.

(131) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure-activity analysis*; Wiley: New York, 1986.

(132) Randi, M. On Characterization on Molecular Branching. *J. Am. Chem. Soc.* **1975,** *97*, 6609-6615.

(133) Kier, L. B. A shape index from molecular graphs. *Quant. Struct. -Act. Relat.* **1985,** *4*, 109-116.

(134) Kier, L. B. Inclusion of symmetry as a shape attribute in kappa-index analysis. *Quant. Struct-Act. Relat.* **1987,** *6*, 8-12.

(135) Kier, L. B.; Hall, L. H. An Electrotopological State Index for Atoms in Molecules. *Pharmaceutical Res.* **1990,** *7*, 801.

(136) Kier, L. B.; Hall, L. H. An Index of Electrotopological State of Atoms in Molecules. *J. Math. Chem.* **1991,** *7*, 229.

(137) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: 1999.

(138) Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992,** *32*, 331-337.

(139) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000,** *40*, 185-194.

(140) Tropsha, A. Recent Trends in Quantitative Structure-Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*; Abraham, D. Ed.; John Wiley & Sons, Inc.: New York, 2003; pp 49-77.

(141) Itskowitz, P.; Tropsha, A. kappa Nearest neighbors QSAR modeling as a variational problem: theory and applications. *J. Chem. Inf. Model.* **2005,** *45*, 777-785.

(142) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design (Methods and Principles in Medicinal Chemistry, Vol 2)*; Waterbeemd, H. v. d. Ed.; Wiley-VCH Verlag GmbH: Weinheim (Germany), 1995; pp 309-318.

(143) PubChem. Structural Clustering. http://pubchem.ncbi.nlm.nih.gov/assay/assaycluster.cgi . 2007.

(144) Jorgensen, W. L.; Tirado-Rives, J. QSAR/QSPR and Proprietary Data. *J Chem. Inf. Model.* **2006,** *46*, 937.

(145) Oprea, T. I.; Tropsha, A.; Faulon, J. L.; Rintoul, M. D. Systems chemical biology. *Nat. Chem. Biol.* **2007,** *3*, 447-450.

(146) Wang, R. X.; Lu, Y. P.; Wang, S. M. Comparative evaluation of 11 scoring functions for molecular docking. *Journal of Medicinal Chemistry* **2003,** *46*, 2287-2303.

(147) Wang, R. X.; Lu, Y. P.; Fang, X. L.; Wang, S. M. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *Journal of Chemical Information and Computer Sciences* **2004,** *44*, 2114-2125.

(148) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954,** *22*, 1420.

(149) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Wang, S. M. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry* **2004,** *47*, 2977-2980.

(150) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Yang, C. Y.; Wang, S. M. The PDBbind database: Methodologies and updates. *Journal of Medicinal Chemistry* **2005,** *48*, 4111-4119.

(151) Community Structural-Activity Resources (CSAR). 8-31-2010.

(152) Parr, R. G.; Von Szentpaly, L.; Liu, S. B. Electrophilicity index. *Journal of the American Chemical Society* **1999,** *121*, 1922-1924.

(153) Wu T-Y. Protein descriptors. 2010.

(154) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual density functional theory. *Chemical Reviews* **2003,** *103*, 1793-1873.

(155) Parr, R. G.; Yang, W. T. Density-Functional Theory of the Electronic-Structure of Molecules. *Annual Review of Physical Chemistry* **1995,** *46*, 701-728.

(156) Agrafiotis, D. K.; Xu, H. F. A self-organizing principle for learning nonlinear manifolds. *Proceedings of the National Academy of Sciences of the United States of America* **2002,** *99*, 15869-15872.

(157) Rassokhin, D. N.; Agrafiotis, D. K. A modified update rule for stochastic proximity embedding. *Journal of Molecular Graphics & Modelling* **2003,** *22*, 133-140.

(158) MathWorks Inc. 2009. MathWorks.

(159) Schneider, G.; Bohm, H. J. Virtual screening and fast automated docking methods. *Drug Discovery Today* **2002,** *7*, 64-70.

(160) Good, A. Structure-based virtual screening protocols. *Curr Opin Drug Discov Devel* **2001,** *4*, 301-307.

(161) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002,** *7*, 1047-1055.

(162) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: Ranking, voting, and consensus scoring. *Journal of Medicinal Chemistry* **2006,** *49*, 1536-1548.

(163) Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V. MedusaScore: An accurate force field-based scoring function for virtual drug screening. *Journal of Chemical Information and Modeling* **2008,** *48*, 1656-1662.

(164) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry* **2006,** *49*, 6789-6801.

(165) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *Journal of Computer-Aided Molecular Design* **2000,** *14*, 731-751.

(166) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003,** *68*, 76-90.

(167) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: Virtual Screening Using Molecular Fields. Application to the DUD Data Set. *Journal of Chemical Information and Modeling* **2008,** *48*, 2108-2117.

(168) Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation using the DUD Data Set. *Journal of Chemical Information and Modeling* **2010,** *50*, 1442-1450.

(169) Kraemer, O.; Hazemann, I.; Podjarny, A. D.; Klebe, G. Virtual screening for inhibitors of human aldose reductase. *Proteins* **2004,** *55*, 814-823.

(170) Ferri, N.; Corsini, A.; Bottino, P.; Clerici, F.; Contini, A. Virtual screening approach for the identification of new Rac1 inhibitors. *J. Med. Chem* **2009,** *52*, 4087-4090.

(171) Muegge, I. Synergies of virtual screening approaches. *Mini. Rev. Med Chem.* **2008,** *8*, 927-933.

(172) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *Journal of Computer-Aided Molecular Design* **2008,** *22*, 169-178.

(173) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D-Biological Crystallography* **2010,** *66*, 12-21.

(174) Liu, Y.; Gray, N. S. Rational design of inhibitors that bind to inactive kinase conformations. *Nature Chemical Biology* **2006,** *2*, 358-364.

(175) Ding, F.; Yin, S.; Dokholyan, N. V. Rapid Flexible Docking Using a Stochastic Rotamer Library of Ligands. *J. Chem. Inf. Model.* **2010**.

(176) Ding, F.; Dokholyan, N. V. Emergence of protein fold families through rational design. *PLoS Comput. Biol.* **2006,** *2*, 725-733.

(177) Nicholls, A. What do we know and when do we know it? *Journal of Computer-Aided Molecular Design* **2008,** *22*, 239-255.

(178) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *Journal of Computer-Aided Molecular Design* **2008,** *22*, 133-139.

(179) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *Journal of Computer-Aided Molecular Design* **2008,** *22*, 179-190.

(180) Jahn A, H. G. F. N. Z. A. Optimal assignment methods for ligand-based virtual screening. *Journal of Cheminformatics* **2009,** *1*, 14.

(181) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *Journal of Computer-Aided Molecular Design* **2008,** *22*, 141-146.

(182) Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate for this Class of Proteins? *Journal of Chemical Information and Modeling* **2009,** *49*, 1568-1580.

(183) Molecular Operating Environment (MOE). 2007.

(184) Noble, M. E.; Endicott, J. A.; Johnson, L. N. Protein kinase inhibitors: insights into drug design from structure. *Science* **2004,** *303*, 1800-1805.